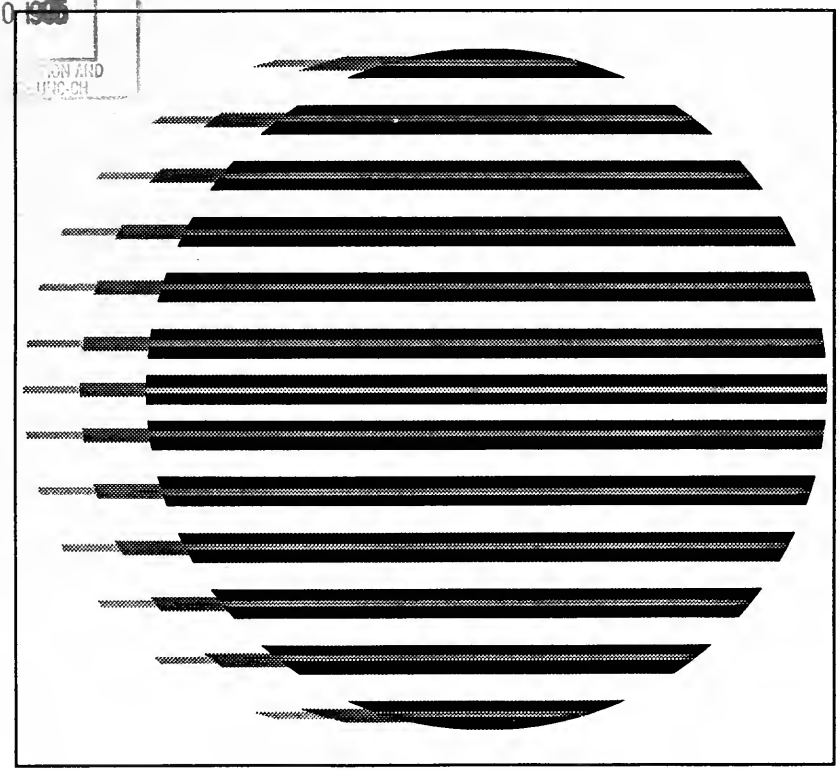
IASSIST

QUARTERLY

VOLUME 19 Spring 1995 NUMBER 1



Printed in the U.S.A.

IASSIST QUARTERLY

The IASSIST QUARTERLY represents an international cooperative effort on the part of individuals managing, operating, or using machine-readable data archives, data libraries, and data services. The QUARTERLY reports on activities related to the production, acquisition, preservation, processing, distribution, and use of machine-readable data carried out by its members and others in the international social science community. Your contributions and suggestions for topics of interest are welcomed. The views set forth by authors of articles contained in this publication are not necessarily those of IASSIST.

Information for Authors

The QUARTERLY is published four times per year. Articles and other information should be typewritten and double-spaced. Each page of the manuscript should be numbered. The first page should contain the article title, author's name, affiliation, address to which correspondence may be sent, and telephone number. Footnotes and bibliographic citations should be consistent in style, preferably following a standard authority such as the University of Chicago press *Manual of Style* or Kate L. Turabian's *Manual for Writers*. Where appropriate, machine-readable data files should be cited with bibliographic citations consistent in style with Dodd, Sue A. "Bibliographic references for numeric social science data files: suggested guidelines". *Journal of the American Society for Information Science 30(2):77-82, March 1979*. If the contribution is an announcement of a conference, training session, or the like, the text should include a mailing address and a telephone number for the director of the event or for the organization sponsoring the event. Book notices and reviews should not exceed two double-spaced pages. Deadlines for submitting articles are six weeks before publication. Manuscripts should be sent in duplicate to the Editor: Laura Bartolo, Libraries & Media Services, Kent State University, Kent, Ohio 44242. (216) 672-3024. Email: LBARTOLO@KENTVM.KENT.EDU. Book reviews should be submitted in duplicate to the Book Review Editor: Daniel Tsang, Main Library, University of California P.O. Box 19557, Irvine, California 92713 USA. (714) 856-4978 E-Mail: DTSANG@ORION.CF.UCI.ED

Title: **Newsletter - International Association for Social Science Information Service and Technology**
ISSN - United States: 0739-1137

CONTENTS

Volume 19 Number1 Spring 1995

FEATURES

Technological change and the provision of documentation for time-series datasets

by Hilary Beedham[1]
ESRC Data Archive
University of Essex,

Acknowledgements.
I would like thank my colleague, Paul Child who supervised the technical aspects of the work to which this paper relates and who contributed generously to the technical content of this paper.

I would also like to thank the U.K. Central Statistical Office, not only for making the Family Expenditure Survey data available to the academic community through the ESRC Data Archive but also for their continuing interest and support of the Archive's work in making the data and documentation more easily available to secondary analysts.

Background.
The Family Expenditure Survey (FES) is a survey of household spending which originated from a recommendation of the Cost of Living Advisory Committee (now the Retail Prices Index Advisory Committee) that such an inquiry should take place as a source for the weighting pattern of the Index of Retail Prices - commonly known as the Retail Price Index (RPI). The first such survey was carried out in 1953/4 and the survey began as a continuous survey in 1957.

The Data Archive at Essex University is now the only and most complete source for these data which are used extensively for secondary analysis by researchers in a varied number of disciplines.

The data collected include not only details of household expenditure such as spending on rent, rates, transport, building maintenance and fuel expenditure but also a substantial amount of demographic information and information collected from a two week diary which is kept by all adults in the household.

The earliest data have either been lost or are not machine-readable. The schedules for the 1953/4 survey are held in their original non-anonymised form at the Public Record Office and are not available for conversion into a machine-readable file. The datasets from 1957 to 1960 inclusive were held on punch cards and are, unfortunately deemed to have been lost. The original data collectors have been generous in support of the Archive's attempts to find the cards but they have never been recovered and further efforts are now thought not to be worthwhile.

The datasets are well documented for almost all of the years for which they are preserved. There is an insuperable problem with the data for the years 1964 to 1967 inclusive as the data layout documents are inadequate for the accurate interpretation of the mixed binary/character data files. The Archive has done extensive work on this over the years but we have been unable to make an accurate interpretation of the files. One researcher with a particular interest has managed to read a few specific variables based on matching consistent codes (such as region) across years but the reliability of even these few variables must remain in doubt. For one of these years there is an additional problem in that only three of the four seasonal quarters has been preserved.

The Problem.
There is nevertheless a considerable amount of data and the FES is available for academic research continuously since 1968. Data are also readily available for the years 1961 to 1963. The type and variety of information included in the survey along with the length of time for which the data are available mean that the FES is one of the most heavily used surveys held in the Archive.

The data have a complex hierarchical structure: they are subject to changes in definitions resulting from, for example, changes in benefits available to the public; methodological changes; changes in coding and content; and changes to the structure of the databases supplied to the Data Archive. The documentation is essential to anyone who wants to use these data and with the support of the depositors the Data Archive strongly recommends that the data should only be used with the full set of documentation for each year in use and no longer offers substantive support to users who fail to purchase the documentation.

This policy is not without problems, however, since although we are funded to provide the data free of charge to most academic users, we do have to charge for the documentation on a cost recovery basis. The total cost of the documentation for the entire set of FES documentation is approximately £775.00 (US$ 1250, approx).

The effects of this on researchers are threefold:

1. Some researchers choose not to use the data because of these costs.

2. Others curtail their research and use only a few years worth of data when they would have preferred to use all.

3. Others insist on ordering all the data but with documentation for only one or two years.

All of these effects are unwelcome and act as a deterrent to good quality academic research despite the use of a very rich source of information.

The Solution.

By 1991 the Archive was becoming increasingly interested in the possibility of providing documentation for on-line use. The provision of FES documentation in this way would mean that the problems associated with the cost and quantity of the documentation to users could be almost eliminated and we began to consider a feasibility study into whether or not it would be possible to convert all of the documentation for the FES into machine-readable files.

Appendix 1 gives an indication of the quantity of documentation which required conversion by information type. The amount varied by year with the earliest years having the least. For 1961, for example, there are only 44 pages which include the data layout, the schedules and a note on the validation tests carried out on the data. In contrast, the most recent dataset (1993) is much more fully documented and has a total of 1324 pages of documentation.

If this could be achieved, we could solve the problems of researchers' access to the documentation by including the files on the same medium as the data, thus charging users only for the medium on which they are sent rather than the copious amounts of paper as at present. This solution also has the potential of reducing staff time required in preparing documentation orders since it removes the need to photocopy the documents as they were required.

There were, however, a number of problems which had to be considered before even a feasibility study could begin:

1. Collation of the documentation.

Key documentation, the schedules and the data layout tables with coding information attached, had always been made available to users via the Archive and was filed in such a way as to make it readily and easily available for both internal and external use. Between 1968 and 1986, other, more comprehensive documentation had been available to users directly from the depositors. In 1990/91, the depositors sent the remaining copies of these 'Information packs' to the Archive for dissemination as required. These were not available for every years and some did not contain all the documentation indicated by their contents pages. Some of these gaps might be filled with the documentation already held in the Archive and some of the missing information could be acquired from the annual reports for the surveys. Another problem was that some of the documents overlapped with the schedules and coding notes we already held. In order to collate the information we had to list the contents of each pack, compare this with the contents list and the contents of each other pack for the same year and then check whether any known gaps could be filled either from the documentation already in the Archive or from the depositors.

2. Quality of the documentation.

As might be expected the quality of the paper and the typeface of the documents varied over the years and both the typeface and the paper quality varied between the earliest and the latest year. Many of the early documents were photocopies of type-written originals whilst later documents are copies of either word-processed files or of output from computer printout. The latter did not all present problems since we can generate equivalent machine-readable files from the data files we hold because they were created using the SIR software. This is not, however, the case for all such documentation because the complex documentation processes in place at the Employment Department (the Government department with responsibility for this aspect of work on the FES) could not all be reproduced with the files we archive. We could not re-create any documentation prior to that for 1986 by this means.

3. Current scanning technology.

We needed to spend a significant amount of time deciding which of the documents could be scanned and which would have to be typed in. From the outset it was agreed that the schedules could not be scanned using existing scanning technology so any feasibility study would have to include provision for typing these in. There were other documents which lay in a 'gray area' with respect to the scanning/typing decision and the only solution was to include comparative time tests for the same documents - scanning versus typing.

4. Which output format should we choose?

A decision also had to be made as to whether the paper should be converted into Optical Character Recognition (OCR) or image files. This decision was relatively easily made on the basis that very large amounts of storage space which would be required if the documentation were to be converted into image files. Also, experience suggested that users would have less problems reading OCR files than image files and we did not wish to embark on such a large project only to create new problems for users.

5. Equipment.

The scanner which was then available to the Archive was a Onescanner attached to an Apple Macintosh II ci computer. The software used was Omnipage. This is a fairly dated set of equipment with a flatbed rather than document feeder and it was clear that we would need a more recent machine if the full project were to be undertaken efficiently. We were aware that any feasibility study using this equipment would necessarily suggest considerably more resources than would be needed if we had a more up to date set-up. However, we were fortunate in learning of a much better scanner in another department and staff there generously allowed us use of the machine during the Summer vacation. There were

restrictions on the use of this but it offered a much better opportunity to produce realistic figures on the resources which would be needed to convert the entire set of documentation into machine-readable files. One important outcome of the project has been to demonstrate the Archive's need for funding for a better scanner.

The feasibility study.
The feasibility study was conducted in a number of stages with some aspects of the work running in parallel.

Stage 1: Preliminary work.
Stage 1 involved both the collection of documentation and the practical testing of the Archive's flatbed scanner. The latter simply demonstrated the impracticality of using such a machine for such a large project. At this stage a list was compiled of what documents were expected to exist for each year (Appendix 1). Having itemised all the documentation, a document typology was developed to serve as a framework in which to apply different OCR settings and to create Recognition Training files (RT files) within the scanning software. Five different types of text were described within this typology, shown on Table 1:

Table 2 gives the estimated proportion of each type of text within the documents.

During this period, after the typology had been created, some of the documents were given to experienced typists in the Archive to compare typing with scanning and determine approximate costs if scanning proved not to be feasible.

Stage 2: assessment.
With the information gained in stage 1 and with access to a Kutzweil 6000 scanner attached to a 486 PC, using Textbridge software, we were able to employ a clerical assistant for a few weeks to take the project forward. During the first week, the documents were carefully collated and many gaps were filled. As so many of the documents proved to be unique, some time was also spent in photocopying the originals: it has long been Archive policy that any extensive internal work is not carried out on original documents as a preservation measure.

Also during the first week, time was spent in familiarising both the clerical assistant and the responsible staff member with the new machine. The learning curve proved to be steep and a number of false starts were made.

There were also problems which were due to the machine being on loan: we were unable to alter basic settings and the software had not been fully installed so that some of the features which would have improved certain aspects of recognition were not available. At one point the system crashed and the machine was reconfigured differently but we were not in a position to rectify this.

Nevertheless, significant progress was made and there were clear advantages to using the more recent equipment:

Table 1: A document typology for use with OCR Scanning.

Typology	Description
Simple text	Text written from left to right of a whole page without non-grammatical breaks. Structural objects such as headings and formatting objects such as indents may be included.
Formatted Text	Includes simple text, lists and ASCII tables i.e. Tables without graphic lines, the columns are separated with spaces or tabs.
Columnar text	Simple or formatted text in a newspaper style layout.
Tables	Columns of information with graphic lines.
Complex Text	Includes simple text, formatted text and graphic lines. Unlike formatted text there may or may not be a relationship between lines of text. A good example of this is schedules.

Table 2: The estimated proportion of each document type contained in each document

Document Typology	Estimated proportions per document
Simple text	20%
Formatted text	33%
Tables	22%
Columnar text	7%
Complex Formatted text	18%

1. Time was saved because of the document feeder;

2. Verification was more efficient since once a correction has been made during scanning, the software 'remembers' it and can apply it elsewhere in the file during further scanning. Verification was slower but more extensive than with Omnipage but the actual scanning was considerably quicker.

3. The Textbridge software allows for zoning where the operator can isolate different parts of a page and apply different settings to each part. This was not fully exploited but is potentially very useful where distinct document types exist on the same page.

4. The software allows the operator to load a dictionary of terms which are specific to the area to which the document relates and also has the option to set lexical and grammatical rules depending on the type of document being scanned. For instance, when tables are being scanned, grammar rules can be turned off.

5. Extensive tests were not run using the available delayed processing facility, partly because of the need to ensure that we did not impinge on the work of the department which loaned us the machine. However, Textbridge allows image files to be created in Tiff format on which normal processing can be undertaken at a later time. Combined with the zoning facility, groups of documents can be processed outside normal working hours resulting in substantial efficiency gains.

Results.
Using the more advanced scanner, the clerical assistant succeeded in scanning the documentation for a full year; deliberately one of the years with the most documentation. This may seem limited, given that he was employed for 3 weeks in total but as has been explained, most of the first week was spent in collating, checking and photocopying documentation. Over a week was spent in familiarisation with the equipment and with running test documents through the system to create the RT files and only a few days of the final week were actually spent on systematic scanning of the documents.

We are confident that the bulk of the FES documentation can be scanned. That which cannot, will have to be typed in and this is expected to be costly because of the complexity of the schedules although it may be worthwhile to examine the possibility of reformatting the schedules so that information is not lost. We would only do this in consultation with the depositors and would not release altered schedules without their approval.

There are two key elements to the successful completion of this project: ready access to a state of the art scanner; and resources to employ a staff member who can make maximum use of this. The clerical assistant who worked on the feasibility study made it quite clear that the scanner was still 'learning' even after completing the work on one set of documentation.

As a result of this work, the Archive has been awarded funds jointly with another ESRC centre at Essex for the purchase of an extremely powerful scanner. It is hoped that resources can now be found for a staff member to work on this and take the project to its conclusion.

References
Family Spending 1993: A report on the 1993 Family Expenditure Survey. Ed. John King. London: HMSO 1994.

Family Expenditure Survey Handbook. WFF Kemsley, RU Redpath & M Holmes. London: HMSO 1980.

1 Paper presented at the IASSIST Conference, Quebec City, May 1995.

APPENDIX 1

Availability matrix of documents within the information packs, year by title.

	Intro	Annex A	Coding Notes	Annex B	Notes on Coll/ dp	Sched's	II & S	V.S	ABTD	cf of codes	Prob's notes
1986		ü	ü	ü	ü	ü	ü	ü		ü	ü
1985	ü		ü		ü	ü	ü	N/A	ü		ü
1984	ü		ü		ü	ü	ü	N/A	ü	ü	ü
1983			ü		ü	ü	ü	N/A	ü	ü	ü
1982	ü	ü	ü	ü	ü	ü	ü	N/A	ü	ü	ü
1981	ü	ü	ü	ü	ü		ü	N/A	ü	ü	
1980	ü		ü			ü	ü	N/A	ü	ü	ü
1979	ü		ü			ü	ü	N/A	ü	ü	ü
1978	ü		ü			ü	ü	N/A	ü	ü	ü
1977	ü		ü			ü	ü	N/A	ü	ü	ü
1976	ü		ü			ü	ü	N/A	ü		
1975	ü		ü			ü	ü	N/A	ü		
1974						ü		N/A	ü		
1973			ü			ü		N/A	ü		
1972			ü			ü		N/A	ü		
1971			ü			ü		N/A	ü		
1970			ü			ü		N/A	ü		
1969			ü			ü		N/A	ü		
1968			ü			ü		N/A	ü		

KEY:

Intro:-	Introduction.
Notes on Coll/dp:-	Notes on collection and data processing.
Sched's:-	Schedules.
II and S:-	Interviewers instructions and sampling.
V.S:-	Variable Schedules.
ABTD:-	Annual base tape documents. (Data layout files with some coding information)
cf of codes:-	Comparison of codes between current and previous years.
Prob's/notes:-	Problems and notes.

Data Rescue: experiences from the Alberta Hail Project

by B. Kochtubajda [1,2]
C. Humphrey [2]
M. Johnson [3]

Abstract
A valuable meteorological data archive collected by the Alberta Research Council over the course of the Hail Studies Project in central Alberta is in jeopardy of becoming unusable as the digital data stored on magnetic tape degrade over time, and expertise in the data collection, calibration, and interpretation becomes scarce. The overall goal of this project was to preserve the digital radar, aircraft, upper air and surface precipitation data along with supporting calibrations and documentation; to transfer this archive to the University of Alberta; and to make the archive available to the scientific community.

There were three distinct operations carried out to ensure the long-term preservation of the archive; retreival of the digital data and all supporting (secondary) datasources; transfer of digital data from magnetic tape to compact disk; and the collection and preparation of relevant documentation describing the data. The archive will provide researchers with a documented dataset to support further research in radar meteorology, climate change, hydrology, cloud physics, mesoscale meteorology and severe weather phenomena.

1. Introduction
The acquisition of atmospheric data is an expensive endeavour, and the data are usually irreplaceable. The subsequent research uses of good data are often not contemplated by the original data collectors. For example, data can be re analyzed to test new hypotheses, or can be used for comparative analyses with other geographic areas. Data may become unusable when supporting documentation is lost or destroyed, or when the physical media on which these data are stored become no longer readable through degradation over time, or through the lack of equipment capable of reading the physical medium due to its obsolete format.

The Alberta Hail Studies Project (1956-1985) was established to study hailstorm physics and dynamics and to design and test means for suppressing hail. Central to these activities was the Alberta Research Council's (ARC) radar facility located at the Red Deer Industrial Airport in central Alberta (Figure 1). A vast amount of data was collected from several platforms to conduct research into precipitation mechanisms, severe storm development, hail suppression, hydrology and microwave propagation.

Since the termination of the Alberta Hail project in 1986, numerous research projects have demonstrated the value of using the Alberta data archive. During the period 1990-1994, 23 archive-based publications have appeared in refereed journals and conference proceedings and 4 scientific reports have been prepared. There have also been nine graduate theses (2 Ph.D., 7 MSc) awarded at 3 universities during this period. The areas of study have included radar meteorology, cloud physics, hydrology/hydrometeorology, computer science, instrumentation, and synoptic, dynamic and mesoscale meteorology. Scientific research and collaborations continue to this day.

Recognition that this valuable meteorological data archive was in jeopardy of becoming unusable as the digital data stored on magnetic tape was degrading over time, and expertise familiar with the data collection and calibration procedures, and their interpretation became scarce, prompted an effort to save this unique dataset.

2. Objectives
The specific objectives of this project were:

1. to transfer the computer-readable radar, aircraft, upper air and surface data from the existing short-term storage medium (magnetic tape) to an archival medium (CD ROM).

2. to collect all available supporting (secondary) data sources and develop the necessary documentation to describe the computer-readable data files.

3. to coordinate these efforts with the University Data Library and develop appropriate mechanisms to make the archive available to the scientific community.

3. Approach and Work Plan
There were three distinct operations carried out to ensure the long-term preservation of the archive; retrieval of the digital data and all supporting (secondary) data sources; transfer of digital data from magnetic tape to compact disk; and the collection and preparation of relevant documentation describing the data.

3.1 The Data Archive
The radar facility near Red Deer consists of a unique polarization-diversity S-band (10 cm) weather radar, a standard C-band (5 cm) weather radar, and an X-band (3 cm) radar used to track aircraft through a transponder

system. With the addition of computer interfaces in 1974, a systematic archive of radar data was initiated. This archive now includes close to 200 Gb of data, representing approximately 12, 000 hours of multi-parameter radar data. In addition to the radar archive, an extensive archive of aircraft, surface precipitation, and upper air data has also been collected. Approximately 18 Gb of data were recorded between 1983 and 1985 aboard an instrumented research aircraft flying through convective storms and cumulus clouds. Also, quantitative precipitation reports (hail and rain) were obtained from approximately 500 ground stations within the radar coverage, between 1974 and 1985 and in 1989. These data exist on several media, including 800, 1600, 6250 bpi magnetic tape and 8 mm cassettes. Comprehensive radar, aircraft and upper-air software packages are also available for data analysis and display.

3.2 Retrieval of the digital data and all supporting data sources

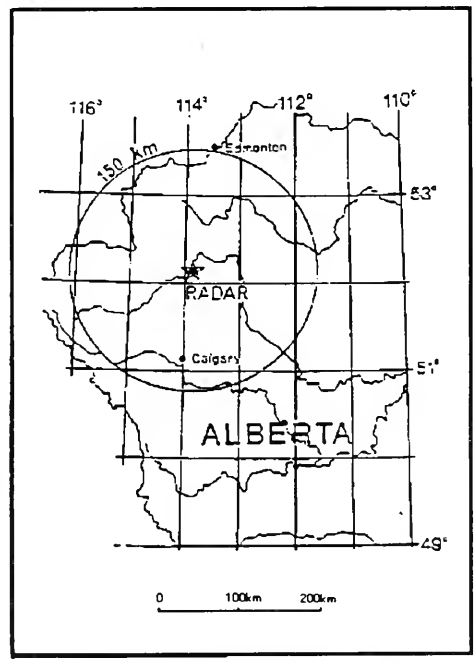
A number of assumptions were made before the digital data were recovered. To provide the broadest range of research potential of the data, unprocessed aircraft and radar data would be provided. This would yield a quicker retrieval of the data and (given the time and budgetary constraints of the project) would result in more of the data being recovered. To maximize the research potential, we would archive all data types and work backward by year, from 1985 towards 1974, thus ensuring a complete multi-platform data set.

3.2.1 Aircraft data

Physical experiments designed to explore the potential of hailfall suppression, and rain augmentation through airborne glaciogenic seeding on convective cells were conducted in central Alberta between 1983-1985, as described in Humphries et al (1986). These studies emphasized in-situ aircraft measurements to investigate natural and artificially modified precipitation processes. The primary observational platform used in these studies was the Intera/ Alberta Research Council cloud physics instrumented research aircraft, a Cessna 441 Conquest, pressurized twin-engine turboprop aircraft. Data from the instruments were managed by a computer based data system which provided data acquisition, recording, and real-time calculations and display (Johnson et al, 1987).

Approximately 300 magnetic tapes were processed for all the research flights conducted from 1983 to 1985. Data were segmented into unique files for each hour of the day. The file names follow the ISO 9660 level 1 standard, and are of the form YYMMDDHH.ADB where YY - year; MM - month (numeric to help in sorting); DD - day; HH - hour of the day; and ADB represents the file extension identifier for Aircraft Data Block. Aircraft data were recorded with Coordinated Universal Time. Yearly index files summarizing the amount of information collected for each hour of the research flight were prepared. The first line provides a brief description of the purpose of each research flight, including the date, start and end time of data collection (hh:mm UTC) and the type of study carried out. The subsequent lines describe the filename, file size (in KBytes and MBytes) as well as the number of records collected (including the 2-D imagery).

figure 1



Location at the Alberta Research Council's weather radar facility and coverage area.

3.2.2 Polarization radar data

The S-band polarization-diversity radar, installed in 1967, operates at 2.88 GHz and has a parabolic reflector antenna with a 6.67 m diameter dish that produces a 1.15_ beamwidth in both azimuth and elevation. The radar sweeps out a helical volume scan rotating at 48_s^{-1} (7.5 s per revolution) and rising 1_ in elevation for every 360_ in azimuth, up to a maximum elevation of 8 or 20_ selected depending on the proximity of the storms to the radar. The radar records data with an approximate azimuthal resolution of 1_, for 147 range gates, from 3 km to a distance of 157 km from the radar, with a range resolution of 1.05 km per range gate.

The radar can transmit any polarization, but has usually transmitted left-hand circular (LHC) polarization with 450 kW peak power. The receiver circuitry digitally records four measurements from the

Table 1: Summary of supporting data sources for 1983 and 1984.

Year	Photo/Maps	Shelf Loc.	Data Timelines	Summaries	Shelf Loc.	Operational Logs	Shelf Loc	Checks	Shelf Loc	Calibrations Data	Shelf Loc.		
1983	Maxmaps	F6 G6	Comp. tape			A6B	Vehicle logs		A5B	Daily oper		A6B	Computer calibs	E5B
	Trans maintenance chart	E5B	Daily timeline			A6B	Daily journals		A5B	Midnight checklist	A6B	Major calibs		A6B
	Computer calib plots	A6B	Radar summaries			E5B F6 G6 A7B B7B	Daily appointments		A5B					
	Cloud photos	D3B	Wx trend sheets			F6 G6	Radar logs		A6B					
			R/A data			C2B D2B E2B	Tape logs		E5B					
							C/S band comparisons		D2A					
1984	Maxmaps	C5B F6 G6	Daily timeline			C5B	Daily appointments		A5B	Transponder checks	B5B	Antenna calibs		B5B
	AMVI disc	C5B	Wx trend sheets			F6 G6	Computer Radar		B5B	Midnight checklist	B5B	Receiver calibs	G2A	B5B
	Cloud photos	D3B	Video flight			F6A	Tape logs		C5B	Receiver checks	B5B	Computer calibs		B5B
	Snow flight photos	E3B	Radar const.			B5B	Radar logs		C5B	Daily oper.	B5B	Major calibs		E5B
			R/A data summaries			F2B G2B A3B				Transmitter checks	C5B			
1985	Computer plots	D5B	Daily timeline			C5B	Appointment diary		A5B	Transmitter checks	C5B	Major calibs		D5B
	CAPPI 'S	F6 G6	Video flight sums		F6B	Daily journal	A5B	Midnight checklist	D5B					
	Major comgraphs	D5B	R/A data summaries			B3B	Vehicle logs		A5B	Transponder checks	D5B			
	Cloud photos	D3B E3B					Tape logs		C5B	Receiver checks	D5B			
	CSU photos	E3B					Radar logs		C5B F2A	Daily Oper checks	D5B			
	Snow trip photos	E3B					Radar calib oper.		D5B					

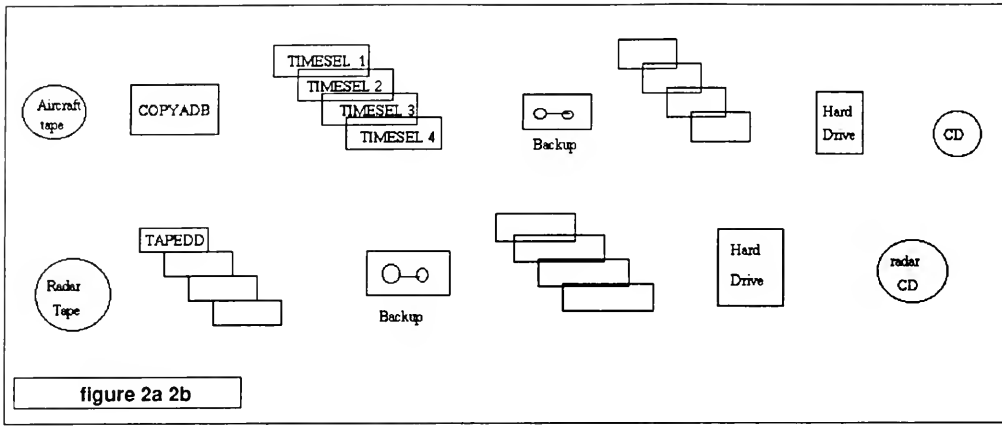


figure 2a 2b

LHC and RHC components from each range bin. These are the RHC co-polar signal power, the LHC cross-polar signal power, and the correlation and phase between the LHC and RHC signals.

Approximately 240 S-band radar tapes (previously copied to 6250bpi) have been examined for the period 1980 to 1985. Radar data were extracted into files containing one complete 3D volume scan, representing either 1.5 minutes for the 8_ scan, or 3 minutes for the 20_ scan. A quality control report was produced with each data file containing information about the azimuth and elevation time histories of the volume scan. The file naming convention adopted for the radar data uses all 11 characters (YYMMDDHHMMR) where YY - year; MM - month (numeric to help in sorting); DD - day; HH - hour of day (24 hour clock); MM - minute of first data in file; and R - represents the data type (S for S-band, C for C-band, Q for quality control reports). Radar data were recorded with local (Mountain Daylight) time.

Table 2: Summary of Data Compact Disks

Data Type	Year	No. Files	No. CDs	Total MB
Aircraft				
	1983	349	5	2947.5
	1984	226	3	1706.8
	1985	166	2	1284.8
Radar				
	1980	13	3	1433.1
	1981	9	1	467.9
	1982	20	3	1657.8
	1983	48	8	3681.5
	1984	56	9	4775.2
	1985	51	11	5093.3

3.2.3 Surface hailfall and rainfall data

Volunteer observations of precipitation events were used to supplement quantitative surface measurements obtained by specially equipped vehicles, which were directed beneath thunderstorms. After each storm, telephone surveys were conducted to collect hail and rain reports. Report cards were also received by mail from volunteer farmers. This information was useful to develop hail climatologies and correlate hailstone characteristics with crop damage.

The surface data collection includes the digital hail and rain report files (YYHAIL.DAT, YYRAIN.DAT), from the telephone surveys for the period 1957 to 1985 (except those files missing from 70-73); selected time-resolved hail and rain truck observations (YYMOBILE.DAT); and daily precipitation measurements collected by approximately 500 volunteer farmers during the months of June, July and August for the period 1975 to 1983

(YYPRECIP.DAT).

3.2.4 Upper air data (LIMEX)

A mesoscale upper-air study, *Li*mestone *M*ountain *Ex*periment (LIMEX-85) was carried out over the foothills and mountains of southwestern Alberta during July, 1985 (Strong, 1989). The objectives of the field experiment were focused on mesoscale convective processes, orographic effects, and interactions with synoptic processes, with particular emphasis on severe storm forecasting applications.

The archive data includes two-hour soundings from nine upper-air sites with an average spacing of 50 km, continuous SODAR profiles, research aircraft soundings at 20-km intervals, surface data from eight automated systems, and an extensive cloud photo set. The compressed upper air data and accompanying analysis software are currently archived on 2 high density diskettes.

3.2.5 Supporting data sources

An equally important component of the retrieval process was the collection of the secondary (supporting) data sources including aircraft mission scientist notes, radar plot summaries, various operational log books, checklists, calibration notes, cloud photographs and video tapes. The data were boxed and transferred from the Alberta Research Council to the meteorology division at the University of Alberta. Subsequently, the boxes were itemized and given a location identifier. The contents of each box were stratified into one of five categories (photos and maps; data summaries and timelines; operational logs, checklists; and calibrations and data. A listing has been prepared which stratifies the data according to the type of information and the year of collection. A subset of the listing (for 1983-1984) reproduced in Table 1, illustrates the variety and richness of the materials available.

3.3 Transfer of digital data from magnetic tape to compact disk

The procedures used to produce the aircraft and radar data compact disks are depicted in Figures 2a and 2b. A series of programs were used to copy unprocessed aircraft data from tape to file (COPYADB), and to generate hourly flight files (TIMESEL). These hourly files were backed up on a series of 8mm EXABYTE data tapes and high capacity SONY compact tapes. Complete 3D volume scan data files and associated quality control reports were generated from the radar tapes (TAPEDD) and also backed up on EXABYTE and SONY tapes.

Compact disks were produced using a Pinnacle Micro RCD-1000 recordable compact disk recorder with Macintosh authoring software and a Quantum 1 GB Fast Scuzzi 3 hard drive for preparing a CD image. The requirements on the hard drive included an average seek time of 12 milliseconds or faster, a transfer rate of 1.2 MB per second or better, and an intelligent calibration feature (ie: thermal recalibration not performed during a continuous read) to avoid a write interruption which would render the CD invalid. The ISO 9660 level 1 standard was selected as the format for the CD-ROMs. This standard allows the same CD to be read and interpreted on Mac, MS-DOS, UNIX, VAX/VMS, and other computer platforms. This includes the restriction of file names to 8 characters, with a 3 character "extension".

An inventory of the compact disks produced is summarized in Table 2. There are 46 CDs in total, including 10 research aircraft data CDs ; 35 CDs containing the S-band polarization data from 197 days between 1980 and 1985; and 1 CD containing the surface precipitation files, dataset documentations, and the radar and aircraft software source codes. Each aircraft CD contains a series of sequential hourly data files and 3 text folders (MAC, DOS, and UNIX) containing the file summaries and a disclaimer. The directory structure of a radar CD is described in Figure 3. A radar CD contains a series of daily files (YYMMDD). Each file contains 4 sub-directories (DATA, QUALITY, CALIB, LOGS). The DATA sub-directory contains the sequence of radar volume scans. The quality control reports for each radar file are located in the QUALITY sub-directory. The calibration text files and the daily radar and transmitter log files are found in the CALIB and LOGS sub-directories, respectively.

3.4 Documentation describing the data

A series of documents have been gathered and/or prepared to describe the various datasets. Aircraft experiment descriptions including study objectives; flight procedures; aircraft description and instrumentation list; 2D image processing; aircraft tag descriptions; daily flight assessments (instrument evaluations); and sensor calibration files accompany the digital aircraft files.

Digital radar and transmitter logs for the period 1977-1985; as well as descriptions of the radar characteristics and scan protocols; data structures; and calibration files accompany the digital radar files. The primary documentation for the surface hail and rain dataset is a coding sheet describing the file format. The daily farmer precipitation files from 1975-1983 has an accompanying text file.

An extensive bibliography of Hail Project related papers has been compiled. The original hard copies are currently stored in the meteorology division at the University of Alberta. Software packages to analyze and display radar, and upper air data, developed at University of Essex and at AES in Saskatoon, have been obtained and can be shared by users. A summary of the archive as it is currently configured is presented in Appendix 1.

4. Conclusions

Potential users of the archive have indicated that the radar and ground measurements dataset would be used to continue severe hailstorm and rainstorm studies; to provide input to

distributed hydrologic models; to carry out radar-based precipitation climatology studies; and to validate numerical models being developed during MAGS, BOREAS, GCIP, or BASE. The aircraft archive would be used to improve our understanding of the chemical composition of cloud water and the processes which affect it, and for icing research. The LIMEX upper air dataset would be used for the atmospheric correction of NOAA-AVHRR data in estimating regional evaporation, as well as in moisture budget estimates and evaluation of evapotranspiration studies.

A documented archive of radar, aircraft, surface and upper air data has been provided from which further research in these areas can be carried out. The retention and preservation of the archives through the University of Alberta will ensure the continued accessibility and long-term survivability of these datasets.

Acknowledgements
This project was financially sponsored by the University of Alberta, Alberta Research Council, and the Atmospheric Environment Service. Mr. S. Kozak and Mr. F. Bergwall assisted in the data retrieval.

The authors also acknowledge the contributions and support of Drs. EP Lozowski, GW Reuter, and T Gan (UoA), Dr. A. Holt (UoEssex), Dr. DR Rogers (Colorado State Univ), Drs. GA Isaac, P. Joe, GS Strong, (AES), and Dr. BL Barge, and Mr. CF Richmond (ARC).

References

Humphries, R.G., M. English, and J.H. Renick, 1986: Weather modification research in Alberta Canada. 10th Conf. Planned and inadvertent weather modification, Arlington, AMS, 357-361.

Johnson, M.R., L.E. Lilie, and B. Kochtubajda, 1987: A data structure for acquisition, analysis, and display of meteorological data. 6th AMS Symp. on Met. Obs. and Instr., New Orleans, AMS, 397-400.

Strong, G.S., 1989: LIMEX-85: 1. Processing of data sets from an Alberta Mesoscale Upper-air Experiment. Climatological Bulletin, 23, 98-118.

1 Paper presented at the IASSIST confernce

2 University of Alberta Edmonton, Alberta 3 Alberta Research Council Edmonton, Alberta

APPENDIX 1: ARCHIVED DATABASE SUMMARY

Data type	Period	Archive filename	Calibration	Documentation
Digital S-band radar	1980-1985	YYMMDDHH.MMR	(n = 1-4)	radar data structure
		YYMMDDnA.TXT		radar characteristics
		RLYYMMDD.TXT		calibration procedure
		TLYYMMDD.TXT	(n = 1-2)	
		YYMMDDnX.TXT		
Digital aircraft data	1983-1985	YYMMDDHH.ADB		Expt desc data index aircraft + instruments daily scores Video logs
surface reports: 1957-1985 hail and rain		YYHAIL.DAT YYRAIN.DAT		hailcard coding form
daily rainfall reports: 500 stations	1975-1983 (June 1 - Sept 1)	YYPRECIP.DAT		file description
mobile reports hail and rain		YYMOBILE.DAT		file description
Upper air data LIMEX-85 9 upper-air stns. 8 auto surface stations	(July 4 - 23, 1985)			
Analysis Software: radar / aircraft / upper air				

Data Rescue: experiences from the Alberta Hail Project

ADDENDUM

The Earth and Atmospheric Sciences Department in collaboration with the University of Alberta Data Library, and the Information Systems Department of the Alberta Research Council have just completed a 15 month effort to rescue the Alberta Hail Studies Project dataset. The project included the organization, retrieval, and formatting of the digital data and all supporting (secondary) data sources; the transfer of digital data from magnetic tape to compact disk; and the collection and preparation of relevant documentation describing the data.

There are 62 CDs in total, including 10 CDs of research aircraft data collected between 1983-1985; 47 CDs containing the S-band polarization data from 287 days from 1979 to 1985, and in 1989 and 1991; 4 CDs of coincident C-band radar data collected on those days when both radars were operating simultaneously (44 days) between 1979 and 1991; and 1 CD containing the surface precipitation files, aircraft transponder files, dataset documentation, and the radar and aircraft software source code. The archive also includes the collection of supporting data (such as; operational log books, manuals, photographs, slides and videos) and 2 diskettes of upper air data and accompanying analysis software.

Example access software (in the C programming language) and documentation has been developed for quick inspection of the original unprocessed aircraft and radar files from the CDs, and as a demonstration of data access.

A set of World Wide Web pages has been developed and is now available on theInternet via browsers such as Netscape and Mosaic. The "Alberta Hail Project Meteorological and Barge-Humphries Radar Archive" can be accessed through network services provided by the Data Library at the University of Alberta, by opening the URL: http:// datalib.library.ualberta.ca/AHParchive/

Data can be accessed in one of three ways. Researchers can obtain hail and rain data files directly from an anonymous FTP site: (ftp://datalib.library.ualberta.ca/AHParchive).

To obtain aircraft and/or radar data from the compact disks, click on the ORDER FORM and submit a specific request. For small amounts of aircraft and/or radar data (e.g.: a single case study), the set of hourly aircraft files or the daily radar directory will be transferred from the CD library and placed in the anonymous FTP site for subsequent retrieval. Requests for larger amounts of data will result in the production of customized CDs and shipment to the researcher for minimal cost.

Use of the Archive, is subject to the following conditions:

1. These data are to be made freely available only to the scientific research community, whether national or international.

2. These data are provided for the exclusive purposes of teaching, academic research and publishing, and/or planning of educational services and may not be used for any other purposes without the explicit written approval, in advance, of the Data Library at the University of Alberta.

3. The Alberta Research Council, the Atmospheric Environment Service and the University of Alberta will be acknowledged in any anticipated presentations and papers associated with the ARCHIVE.

4. The citation to be used for the ARCHIVE is: Alberta Research Council. The Alberta Hail Project Meteorological and Barge-Humphries Radar Archive: [computer files], Edmonton, Alberta, CANADA. Alberta Research Council [producer], University of Alberta Data Library [distributor]. August 1995.

<URL: http://datalib.library.ualberta.ca/AHParchive> <URL: ftp://datalib.library.ualberta.ca/AHParchive

This project was financially sponsored by the University of Alberta, Alberta Research Council, and the Atmospheric Environment Service.Thank you for your support of this project.

Bob Kochtubajda Ed Lozowski Chuck Humphrey Steve Kozak Mark Johnson Ford Bergwall

A Functional Approach to Documentation and Metadata

by Stephan Greene[1]
University of Maryland,
College Park

Introduction

The continuing struggle for documentation standards for social science data is reflected not only in the purely archival context, but also in the context of the varied research that makes active use of archived data. The documentation of archived social science data is, of course, a central component in archival practice. Documentation serves to describe archived data for those who gather data in archives, and functions as an aid to the cataloging, and subsequent retrieval and dissemination, of data. For social science researchers, the end users of data, documentation functions as the primary vehicle for gaining an understanding of the nature of raw data. This understanding is critical if researchers are to be successful in their secondary analyses of the data.

The work discussed in this paper is motivated by the desire for documentation to remain useful as the data it describes is transformed through analytical procedures. It addresses some traditionally ignored user-based issues of data manipulation and integrity, how those issues imply structural weaknesses in methods of data documentation, and how the consideration of these unmet requirements can guide the design of improved documentation. This functional approach, emphasizing considerations of practical data use, leads to a rethinking of traditional documentation from that of a generally static reference document to a dynamic entity more appropriately considered a form of metadata. The correct provision and understanding of metadata, any information that adds meaning to the base data (McCarthy 1982), is increasingly susceptible to compromise as the base data move from static printed tables, as often seen in codebooks, to the dynamic environment of the interactive exploration and analysis tools now appearing on computers in a variety of settings. It is within the context of this dynamic functionality that the work described here is most relevant.

In this paper, I will describe my preliminary implementation of the use of metadata processing in data derivation in an electronic historical atlas of demographic data. The implementation experience has suggested useful concepts regarding documentation and metadata. In addition, I will describe a system for producing machine-readable documentation for social science databases developed at the Swedish Social Science Data Service at Göteborg University . This system acts in many ways as an archival response to some of the extant problems of documentation, as it seeks to address both the functional requirements of data users and their varied software tools, and the archival needs of an archival agency. Finally, I will discuss some approaches I have been exploring for further addressing the problems of documentation targeted in this research. The disciplines and their aspects that may be useful are primarily techniques of knowledge representation in artificial intelligence, elements of traditional library science, and considerations of data theory and dimensional analysis. The eventual goal is a formal model of social science data and its analysis which can directly guide the design of documentation and metadata.

A Preliminary Implementation

As the use of micro computers continues to increase, so does the number of computer-aided studies of numeric social science data. In addition to standard commercial spreadsheet, database, and statistical packages, unique data management and analysis applications tailored to specific purposes, which might generically be called demographic or social science data information systems, are beginning to emerge (Miller and Modell 1988). Any thoughtful consideration of the nature of computer-aided studies of social science data must acknowledge that these environments are permeated with numerous opportunities for misinterpretation and incorrect manipulation and analysis of data (Conroy 1994; Greenstein 1994).

The humanities and social science research communities have identified the need to proceed with caution in computer-aided data analysis (Greenstein 1994). To date, these researchers have only their own expert knowledge to guide them in the rational manipulation of data. Yet even the most skilled expert is subject to generating erroneous data, simply by making a small typographical error resulting in an incorrect variable reference. Moreover, the use and potential abuse of social science data now extends beyond the research establishment. Social science data is now accessed in forums such as dial-up public access Internet sites and other consumer-oriented on-line services, which provide simplified access on inexpensive machines (Conroy 1994). The potential for uninformed use of data is thus amplified. The underlying structural reasons for the pitfalls of data use deserve increased attention.

One specific area which threatens the integrity of social science data is in the derivation of new variables from the existing variables of a social science database. Such procedures are typically performed with statistical packages.

The Great American History Machine (GAHM), an interactive historical atlas, is one specialized software product that exhibits the problems associated with data derivation. GAHM provides a browser with 200 years of United States census and election return data at the county level. The program supports the arithmetic combination of basic count variables such as census population counts into derived variables such as rates. Basic descriptive statistics and a choropleth map can be displayed for each basic or derived variable (Miller and Modell 1988). In this particular application, as in many others, it is entirely the user's responsibility to be sure that the derivations he or she performs are correct. Typical errors include dividing one rate by another, or adding a monetary value expressed in thousands of dollars to a monetary value expressed in millions of dollars without an equalizing conversion. The possibilities for error are as numerous as the possibilities for the derivation of new and interesting variables.

To begin to solve this problem, a preliminary assessment of common data types and attributes was gleaned from a selected sample of census and supplemental data sets resident in the GAHM database. A prototype facility for the support of unit control and error checking in the derivation of new variables from existing variables was also implemented. Metadata in support of unit control was identified and entered for a subset of the GAHM database. This enhanced data subset was then used to develop and test the prototype. The data structures and procedures for handling variables in the C-language code for GAHM were augmented to incorporate the new metadata. Similarly, the code that implements GAHM's existing abilities for data derivation was augmented to process the metadata.

GAHM's data derivation is accomplished by an expression evaluator. Through a simple point-and-click interface, users can combine variables from the database within an arbitrary algebraic expression. Like a simple calculator, the expressions may utilize addition, subtraction, multiplication, division, exponentiation, unary minus, and several additional functions like square root. It is the generality and ease of use of this expression facility that make it at once powerful and problematic. It is, unfortunately, quite easy for the user to commit errors in unit matching when combining variables algebraically. For example, users often may sum variables expressed in entirely different units. For a metadata structure to be useful for identifying mismatched units in syntactically correct expressions, it seems intuitive that similar generality is required of the metadata processing used to ensure semantic correctness. Thus, the method by which generality is achieved in basic data derivation was used as a model for the implementation of the use of metadata in data derivation.

The existing GAHM expression evaluator is implemented as a sort of mini-programming language within the program. Internally, the program uses a data structure known as a last-in-first-out queue, or a stack (Wulf, et. al. 1981), for the manipulation of the operands in an expression. The operands on the stack consist of constants, database variable values, and intermediate values occurring as the expression is computed. Metadata was incorporated into this scheme by creating a second, parallel stack for metadata data structures. Thus, in compiling a data stack to represent an expression, general "compile-time" checking of *syntactic* correctness of the expression can be performed. As the evaluation of syntactically correct expressions is carried out, the metadata stack provides for general "run-time" checking of *semantic* correctness.

Under this general scheme, each variable in the test database was specified with the following items of metadata: scale, unit type, unit, and weight. Options for the scale field were ratio, interval, ordinal, and nominal (Stevens 1946). The initial implementation of the use of this metadata in data derivation concentrated on the four basic operations: addition, subtraction, multiplication, and division. In cases where meta-operation routines for checking semantic correctness detect a possible error, a warning is displayed to the user describing the possible source of a problem. The user is given the option of aborting the procedure, or proceeding with the calculation despite the possible error. If a (potentially) erroneous calculation is carried out, the program tags the resulting unit as "not determined." The existing GAHM interface was modified slightly to include display of the resultant unit value to the user as part of its data description display. Greene (1994) discusses the implementation, as well as some relevant technical database issues, in much greater detail.

Results

There are two primary results from this experimental implementation. First, rudimentary support for some common derivations in social science data manipulation was achieved. Derived units were automatically and correctly expressed to the user. Nominal data was treated correctly for this first time in the GAHM application. Second, from a programming viewpoint, the use of a metadata stack parallel to that of the base data stack proved to be an elegant mechanism for providing general semantic analysis of the data.

Most important, however, are the broader implications of these results. The solution, as implemented thus far, suggests an approach that would address the broad issue of managing social science data in a dynamic environment: *metadata must closely follow the data it describes through any transformational procedure* . Metadata must *guide* the application of transformational procedures, and then *describe* the transformed data. It must undergo parallel procedures, tightly coupled with the procedures applied to the base data, that result in transformed metadata that both reflects the transformation of the base data and provides guidance for further transformations of the data.

The principle of integrated metadata has direct implications for social science database management. Social science databases, once compiled, consist of primarily static data. Its associated metadata is also essentially static. If analytic tools are to offer dynamic data manipulation facilities, they must do so for both types of data. Plans for the design and development of software tools that aim to improve semantic support through the use of metadata will benefit greatly by considering both metadata and metadata processing as integral to all data manipulation functions. It should no longer be sufficient to provide information as to the contents of a social science database with a printed document, or even a machine-readable version of that document. This information should be coupled with the database as a structured metadatabase that can be processed in tandem. An integrated, structured metadatabase is dynamic documentation. It is more than an on-line codebook, and it is more than a context-sensitive help system. It is a knowledgeable link between data and the software manipulation of data.

Integrated metadata could and should be used to trace the history of a data analysis cycle. Users should be able to trace links to their original data long after they have been working with measures derived from it. Providing these functionalities should contribute to the development of a "metadata culture" in which data referenced in an apparent vacuum is unacceptable. While users should always be able to override the suggestions of any error checking mechanism in data manipulation, as they can in the experimental case just described, the fact that they must explicitly do so may help force more rigorous defenses of these deviations from mainstream methodology. Integrated metadata can help prevent unintentional errors, and perhaps it can also improve the climate in which decision making from social science data takes place.

There are many limitations, however, to the experimental solution done in GAHM. Additional work lies in the development of a more complete treatment of error checking in data analysis. A litany of concerns must still be accounted for, as described in Greene (1994). The logic of semantic analysis of data and data manipulation is complex, and a more theoretically grounded approach is needed for a comprehensive solution. Archival strategies, then, must remain flexible. It is not yet possible to commit to any one approach to documentation, as data users are currently performing a wide variety of tasks related to social science data.

An Archival Response

The Swedish Social Science Data Service in Göteborg, Sweden, has developed a documentation system called A-Side (Archival System for Interoperable Data Exchange). This system, a UNIX application written in C and using the X11 Window System, produces a family of machine-readable, variable-level documentation formats. The initial process of generating machine-readable documentation with A-Side can be quite resource intensive if no electronic text is initially available. However, once the data is entered, many possibilities arise.

A-Side's primary output resembles a traditional OSIRIS codebook. Introductory study information is followed by variable descriptions along with residuals and other detailed information pertaining to each of the variables and its supporting raw data. The OSIRIS output file is the archival format that no user will ever see. It is a highly structured, or tagged, ASCII-only (and thus neutral) format. Subject to the life of the storage medium used, and the survival of knowledge of ASCII, the format and meaning of these files can always be understood with careful study, even without the benefit of knowledge of OSIRIS formats. In this context, the rigid structure of the old OSIRIS format is an asset.

More importantly, the format is structured enough to allow for the relatively easy authoring of small utility programs to generate other formats. The A-Side system has recently integrated a number of these utilities and can now, with the invocation of a single menu option, produce HTML codebooks, SPSS setups, as well as several other rich-text formats for printing and on-screen display. The production of SGML formatted files, or files in any other format not yet conceived, should be relatively easy to implement, given the current status of the system. Internally, the A-Side system maintains a great deal of variable-level metadata, and other, richer, primary output formats can conceivably be generated by it. Thus the family of supportable formats is open-ended.

The approach of the A-Side system is characterized by *interoperability* and *sustainability*. Serving diverse needs requires the ability to interchange data and easily generate formats for various purposes. The system is currently positioned to serve a diverse user community, and should scale up well to serve future needs yet to be defined. It is sustainable in that it will always generate a neutral archival format that can be easily adapted to new software and new formats. The system is best viewed as a means to an end, or rather, to many ends, though it does indeed perform some useful core functions for generating documentation. But its most compelling feature is its ability to facilitate the use of data and metadata with other software for more substantive purposes.

Formal Documentation Design

Keeping in mind the idea of integrated metadata, while remaining positioned for whatever formatting requirements the future might bring, we can begin to think about what will improve documentation, and how we can design these improvements. For documentation to become a structured, integrated, and dynamic complement to data, it must be formalized. I believe there are several approaches that may prove to be useful in this effort. The first is that of enumerative taxonomy in the traditional sense of library

science. The description of social science data can benefit from an enumeration of the kinds of things such data addresses. Some form of "off-the-shelf" classification should be available to data documenters. This would include some elements of authority control, such that we might begin to see easier data interchange and integration. The provision of a comprehensive characterization of data to which documenters might appeal can ease the process of documentation, and might help support the increased generation of documentation by primary investigators themselves. Geo-spatial data management is ahead of social science data management with respect to authority control and data interchange. Geo-spatial data managers may appeal to entities such as place-name authorities, and data interchange standards are already established.

Second, the realization of functional, operational metadata as documentation can benefit from current work in the development of ontologies for knowledge sharing, a research movement in knowledge representation in artificial intelligence. As used in artificial intelligence, an ontology is a formalized declaration of domain contents. Unlike taxonomic approaches, an ontology specifies domain contents with a *canonical basis* (Sowa 1984), which structures the content more deeply by constraining the types participating in different relationships, in effect creating a concept system rather than merely a list of instances and attributes. An envisioned ontology of social science data would formalize descriptive concepts relating to the entities typically described by social science data, as well as the more interpretive concepts of data and measurement theory, scaling techniques, statistical methods, and dimensional analysis.

A particularly useful example of an ontology is the Engineering Mathematics ontology (called EngMath) developed by Gruber and Olsen (1994). This ontology supports modeling of an engineering perspective of the physical world. This ontology can serve in some respects as a substrate for an ontology of social science data, which will exhibit some mathematical similarity to engineering. Mathematics, as used by both engineers and social scientists, is intended to represent quantitative phenomena. Given that, many of the specific design elements of EngMath are directly applicable to an envisioned ontology for social science data. EngMath formalizes, among other things, conceptualizations of *physical quantity*, *physical dimension*, and *units of measure*. The fact that these three concepts are separated is the key design innovation of this ontology. Physical quantities attempt to represent quantifiable aspects of the real world. Briefly put, "quantifiable" means that the measured entities "admit of degrees" (Ellis 1966) rather than being a yes-or-no attribute, in a qualitative sense. The ability to be quantified also implies the ability to be algebraically manipulated.

The separation of physical quantities from the units of measure used in their expression is a useful abstraction in that physical quantities are fundamental notions, while units of measurement are merely matters of convention. Distinguishing the conceptualization of units of measure from the conceptualization of physical quantities allows for the expression of physical quantities without committing to a particular system of units, of which there are several in common use. More importantly, the distinction between the two concepts supports the straightforward conversion from one conventional system of units to another, which in turn supports the comparison of similar physical quantities that are expressed with different units of measure.

Dimensional analysis, developed as a technique for analyzing the behavior of physical systems, informs the evaluation of mathematical operations on physical quantities. The separation of dimensions from quantities in the engineering mathematics ontology supports direct application of the principles of dimensional analysis. That physical quantities are characterized by physical dimensions is what distinguishes them from abstract numeric entities. The distinct conceptualization of physical dimensions provides many advantages. Useful algebraic or comparative operations on physical quantities must exhibit *dimensional homogeneity*. For example, it is meaningless to compare a measure of mass against a measure of length, or to add a measure of time to a measure of temperature. The distinct notion of physical dimension allows the enforcement of dimensional constraints independent of particular instances of physical quantities or units of measure.

EngMath provides a good example of a formalization of domain contents. A unit conversion program has been written, using EngMath, to facilitate the interchange of data about physical quantities in engineering. EngMath thus helps solve problems with engineering data that are similar to the problems of managing the manipulation of social science data. There are limits, however, to the degree to which mathematics, as formalized by EngMath, will support the mathematics of social science data. The limits appear primarily in the area of dimensionality. Engineering mathematics enjoys general agreement among its users with respect to basic dimensions such as mass, time, length, and temperature. More complex notions of dimensions, such as force, are expressed in terms of the basic dimensions. In social research, there is little agreement on what to measure, and the dimensions of humans and their artifacts and activities are less well defined. Within the formal abstract algebra used in the engineering ontology, important social science quantities, such as persons or dollars, have as their dimensions the "identity dimension", or in the terminology of dimensional analysis, they are *dimensionless*. Research in human geography and social science data theory (Haynes 1975; Jacoby 1991) has identified the need for further investigation into the current limits of dimensional understanding. Dimensionless count data constitute a significant proportion of available social science data.

Beyond the specific example of EngMath, emerging design principles for ontologies (Gruber 1993) specifically meant for knowledge sharing dovetail nicely with the requirements of structured metadata as documentation for social science data. The first design principle is that of *clarity*. The formalism required in ontology design will enforce clarity in the definitions of the concepts of social science data. Greater clarity and specificity in the collection and dissemination of social science data will always be welcomed.

The principles of *monotonic extendibility* and *compartmentalization* are related guidelines for ontology design. An understanding of the purposes for which a conceptualization will be used should inform the design process. Users of an ontology should be able to extend it for their own purposes monotonically, that is, without requiring changes to the base ontology. Compartmentalization supports the monotonic approach to extendibility. Where it is possible, an ontology should be broken down into component ontologies. This allows users to select those components that are useful, without being forced to inherit those that are not. As users extend ontologies for their own purposes, their extensions should likewise be compartmentalized. Other users may then access extensions, with the same benefits. Given the diversity of methods in data collection and analysis in social science research, an approach that emphasizes a basic core conceptualization that is agreeable to most potential users, and that can be extended as needed without alteration, is inherently appealing.

Another design principle is *minimal encoding bias*. This principle can also be thought of as the *parameterization of convention* (Gruber and Olsen 1994). Many descriptive elements, such as units of measurement or natural language titles, are merely matters of convention, rather than basic concepts. The use of conventional terms is discouraged. This design principle helps support the goal of inter-agent communication and knowledge sharing, as it encourages conceptualization of core concepts independent of the conventions typically used to describe them. This applies directly to the needs of social science data, where data sets collected by different agencies differ radically in their conventions, such as their methods of naming variables, or the systems of units used to report data. Systems of measurement, which are conventions, must be parameterized. Doing so will support efforts toward data integration and interchange.

Another principle of ontology design is *minimal ontological commitment*. This principle stresses designing the weakest conceptualization possible to support the purposes for which it is designed. For example, to the extent that it is possible, formalized descriptions of data should exclude interpretive elements. Doing so maximizes the number of researchers that will agree to use the descriptions, as such descriptions would not require commitment to particular interpretations. The principle of minimizing commitment works hand in hand with compartmentalization, which also reduces the degree of commitment required by any single ontology by encouraging small, modular ontologies. The minimization of ontological commitment may begin to address some of the structural problems that have plagued the search for generally acceptable methods to describe social science data. Controversies abound over the design, collection, interpretation, integration, and analysis of social science data. To the extent that it is possible, an ontology should support data interchange without requiring absolute methodological harmony.

Conclusion

Modern, dynamic data access and manipulation presents new challenges for the processes of collecting, describing, disseminating and analyzing social science data. The experimental solution developed in the context of GAHM shows promise, and suggests some broader principles that may guide the design of data description in a dynamic environment. Archival documentation strategies must be flexible and adaptable, and the A-Side system attempts to meet current needs, while remaining positioned to meet future needs once they become more clear. Documentation standards are difficult to define given the diversity to be found in all aspects of social science data management. Forms of documentation and methods of documentation production that support interoperability and interchange of information will provide the most benefit in the long run.

The design of new forms of documentation that begin to meet some of the concerns of data description outlined in this paper can benefit from considerations of taxonomy and ontology. The goal is to formalize, as much as possible, the contents of the domains of social science data. The resulting data models are then available to guide the generation of documentation. The design of formal conceptualizations of social science data can begin with a comparison to the treatment of engineering mathematics. Such a comparison, guided additionally by consideration of social science data theory, as well as dimensional analysis as applied to social science data, will help to isolate and eliminate weaknesses in the handling of social science data and lead to formalized ontologies, and thus similarly formalized metadata and documentation, for social science data.

References

Conroy, Cathryn. 1994. People, by the numbers. *CompuServe Magazine*. October, 32-36.

Ellis, B. 1966. *Basic Concepts of Measurement*. London: Cambridge University.

Greene, Stephan. 1994. Metadata for Social Science Data Derivation: A Preliminary Approach. Manuscript.

Greenstein, Daniel I. 1994. *A Historian's Guide to*

Computing. Oxford: Oxford University Press.

Gruber, Thomas R. 1993. Toward principles for the design of ontologies used for knowledge sharing. Technical report KSL-93-04, Knowledge Systems Laboratory, Stanford University.

Gruber, Thomas R. and Gregory R. Olsen. 1994. An ontology for engineering mathematics. Technical report KSL-94-18, Knowledge Systems Laboratory, Stanford University.

Haynes, Robin M. 1975. Dimensional analysis: some applications in human geography. *Geographical Analysis* 7: 51-67.

Jacoby, William G. 1991. *Data Theory and Dimensional Analysis*. Newbury Park, CA: Sage.

McCarthy, J. L. 1982. Metadata management for large statistical databases. In *Proceedings of the Eighth International Conference on Very Large Databases*. Saratoga, CA: VLDB Endowment.

Miller, David W. and John Modell. 1988. Teaching United States history with the Great American History Machine. *Historical Methods* 21(3):121-134.

Sowa, John F. 1984. *Conceptual Structures: Information Processing in Mind and Machine*. Reading, MA: Addison-Wesley.

Stevens, S. S. 1946. On the Theory of Scales of Measurement. *Science* 103(2684):677-680.

Wulf, William A. et al. 1981. *Fundamental Structures of Computer Science*. Reading, MA: Addison-Wesley.

All scale types except ordinal actually occurred in the test database.

Documentation - what we have and what we want: Report of an enquete of data archives and their staff

by Karsten Boye Rasmussen[1]
Danish Data Archives

Abstract:
A report from an enquete. Data professionals have given their views to questions on a future codebook format ("should it be SGML?", "should it be supported by vendors?", etc.). This forms a description of "what we want". On the other hand archives have described their holdings with regard to levels of machine-readable documentation. This section focuses on the present and on the actual data thus: "what we have".

The paper presents the key figures from questionnaires sent out by "The IASSIST Codebook Action Group". Maybe we are moving in a direction demanding less knowledge about special formats from our users? Hinting to the conference theme the subtitle could be "Access for other than Partners?".

Background Information - History
At the IASSIST Conference in Edinburgh in May 1993 several sessions were centred around documentation, and some specifically with documentation at the variable level (codebooks) as opposed to documentation at the study level (study descriptions). The sessions concerned with codebooks were: "Roundtable on Codebooks" (Wednesday, May 12) "Poster session on production of codebooks" (Thursday, May 13), and the "Codebook session" (Friday, May 14). I chaired the first and last of these sessions and the DDA contributed to the "Poster Session". The willingness to present papers at the "Codebook Session" as well as the many arguments and the eagerness of discussions all pointed out that many professionals were interested in working in the area of codebook documentation. The papers[0] at the "Codebook Session" contributed to many discussions at the conference, and after the conference the paper by the ICPSR director Richard Rockwell created many e-mail discussions.

At the business meeting at the conference an action group was formed and named: "Codebook Documentation of Social Science Data". I was appointed the chairman or co-ordinator of this action group and during 1993 was joined by Lennart Brantgärde and Bill Bradley as formal members of the action group.

My intentions with the action group was broadcast - on the IASSIST listserver in late May 1993 - as follows:

During the last years there has been some confusion concerning who are making which changes to the OSIRIS codebook. The OSIRIS codebook format has for more than twenty years been the de-facto standard for archives around the world. The most obvious reasons for this standard are that the OSIRIS codebooks can store full text, are input to retrieval systems, and that the codebooks are easily converted to other formats (SAS and SPSS). I propose that the task of the working group is to remedy this confusion by:

1) Identifying the tasks and areas that cannot be handled by the OSIRIS codebook in its present form. Two examples: A) Many archives are looking for a feature of presenting tabulations in the codebooks - not just frequencies. B) A less rigid print out of codebooks not limited by the original card image input format. It is important to note that the first is a problem of storing a type of information in the codebook which does not fit the present format. The second concerns the utilisation of the codebook, and could be changed without changing the codebook format (by flowing the text, and using a different font).

2) Identifying the number of data sets at archives all over the world. The data sets should be grouped by the level of documentation: A) full text machine readable codebooks (what format?); B) abbreviated machine readable documentation (OSIRIS dictionary / SAS / SPSS ?); C) no machine readable documentation (paper / scanned information). Data sets that are originally stored at other archives (ICPSR etc.) are to be counted only at the original archive. Another question would be whether the archive is producing machine readable documentation at all.

The rationale behind the second identification is that archives who have produced and stored machine readable documentation (e.g. OSIRIS codebooks) will be able to convert these to the new format (missing the special features of the

new format). There will not be invented a new codebook format which automatically documents what has not been documented. The archives who now uses the same documentation format will be able to share software for making the conversion to the new format.

I must emphasize, that it is not my intention that the codebook working group should produce a complete proposal for "This is how a Codebook should look like". We have no way to enforce a new standard other than waiting for people and organisations to realise its superiority to older standards. The task of the working group is to identify and structure the problems: these are common problems; these are problems with historical data; these are problems with complex data; etc.I think one feature of the new codebook format can be revealed now: It has to be so flexible that new features do not require a new format.

Another reason for not putting forth a new codebook standard is that it is my belief that documentation at both the study description level and the codebook level (and levels in between) has to be integrated.

My plans for the actions group were not fully met. One of the things that things that changed was - not surprisingly - the time schedule. I had planned to report to the IASSIST conference in 1994, but due to other obligations this was postponed till 1995.

In the meantime several activities took place. In Europe a CESSDA seminar with the title "Variable Level Documentation" took place at the SSD in Göteborg in August 1993[1]; the participants list was not restricted to Europeans. In August 1994 a CESSDA seminar was held in Grenoble[2], this time the theme was "Networking and Internet?", again ICPSR willingly sent a participant so the perspective was broadened and more global than just European. Although the term "codebook" was not part of the seminar title the searching and availability of codebooks on Internet - and therefore also the format of codebooks -were discussed intensively during the seminar. In October 1994 ICPSR announced a commitment in the development of an SGML DTD for codebooks. This issue was addressed again in mid February 1995 when ICPSR announced an international committee.

The questionnaires

In the summer of 1994 I formulated two questionnaires. One questionnaire was individual and gave the opportunity to present personal views on present and future codebook formats. The other questionnaire intended to count the number of studies at different archives and to show the distribution of different levels of documentation. In late autumn I received comments and advices concerning the questionnaire from a group consisting of Charles K. Humphrey (Data Library of University of Alberta), Lennart Brantgärde (Swedish Social Science Data Archive in Göteborg), and William Bradley (Canadian Health and Welfare, Ottawa).

The population and the return rate

Finally on 29 November 1994 the two questionnaires were sent to the listservers consisting of the IASSIST membership (178 recipients), Official Representatives of the ICPSR (195 recipients), and to the small IFDO list (27 recipients). A number of individuals were on more than one of these lists. The guess is that individuals from around 200 institutions were given the opportunity to answer the questionnaires. All recipients had furthermore the opportunity to forward a copy of the questionnaires to other individuals whom might be interested.

A pull for the return of the questionnaires were made shortly after the announced deadline on January 15th 1995. Several questionnaires were received thereafter and the last questionnaire was received at the end of February. At that time the number of received individual questionnaires had reached 50, and the number of institutional questionnaires were totalling 20.

These numbers do not hold evidence of the representativeness or the missing representativeness of the collected data. The investigation was never intended to be representative. However it is a fair assumption that the individuals that took the time to answer the questionnaire are individuals that have an interest in the development of documentation formats. And it is a fact that the archives that have returned the institutional questionnaire are principally amongst the national social science archives. All though it is disappointing that not all IFDO members made the effort of answering the institutional questionnaire.

The individual questionnaire

As shown in the "Appendix 1" the individual questionnaire is mainly about 25 different assertions about codebook documentation that individuals rate with their level of agreement on a five level scale from "strongly agree" to "strongly disagree". The middle category was labelled "indifferent, don't know" and this presented a problem to some individuals as these two answers are not completely identical.

The methodology of exposing individuals to a battery of items is well know. Looking back a feature that limited the amount of points which each individual could distribute would have been effective. It is much easier to answer that a lot of factors are important than actual to rank the factors and point out that "these are most important". On the other hand the distribution via e-mail called for both a very simple layout and a simple question structure. This led to the concentration on the battery of items without any filtering structure or deepening sub-questions.

In "Appendix 2" the distribution of the answers in these 25 variables is shown together with information about the mean and the number of non-missing answers. The mean was simply calculated by appointing the values 1 through 5 to the five answer categories:

1 strongly agree
2 agree
3 indifferent, don't know
4 disagree
5 strongly disagree

The treatment of the items - calculating the mean - implies that the items are viewed as belonging to an interval scale. Lots of arguments can be put forth in favour of or against this decision. In this context I find that "keep it simple" will suffice as legitimisation of this manoeuvre, as a more appropriate measure like "mode" will lose information. If you are interested in viewing the actual distribution of the answer categories for each of the items you should take a look at "Appendix 2".

As the direction of the assertions differs a high mean on one variable and a low mean on another does not necessarily imply that these two variables do not support each other. The mean can range from 1 to 5. In order to compare two variables of different direction you should keep in mind that a mean of for instance 4.2 is equally strong as a mean of 1.8. Means around the number 3 implies that the community has no fixed strong feelings about this particular assertion.

In the following the themes of the 25 assertions have been boiled down to a few headings.

The need for standards

I do not intend to enter a long philosophical discussion about standards. I am sure that we are all aware of the benefits of standards. It is equally true that we all spend time getting from one standard to another, e.g. getting from one analysis package to another. The common sense meaning of standard in this context is "as a rule". We expect to be able to connect our electronic equipment when we move to a new house, the plugs are supposed to be "standard". But having been to IASSIST conferences we know that this is not true when moving between countries. Thus the less common sense and the more sophisticated the more standards are available. Or to put it more precise: the many standards are a painful fact of life.

So it is no surprise that the lead question among the assertions - "1. There is no great need for standardization of codebooks" - receives a very strong disagreement value (mean 4.2). There is a need for standards, and we can continue the search for the standards. The ultimate alternative to the codebook - namely "no codebook" - is considered a very bad solution. "2. A data user should be content with a study description and photocopies of relevant pages from the questionnaire", with a mean of 4.3 this item receives the strongest disagreement of all 25 items.

One of the currently used and widespread standards is OSIRIS[3] and this format is drawn into the discussion in "5. There is a great need for more structured information than is available in the OSIRIS codebook format" that shows a weak agreement (mean 2.6). Half the individuals answer "indifferent, don't know" and maybe they answer the latter simply because they don't know the OSIRIS format and its structure. Other formats - and these are more commercial and more available formats - are also given the same weak agreement (mean 2.6) in "10. Let us stick with commercially supported formats for social science data (e.g. SAS and SPSS)", but it should be noted from the distribution that there is a higher variance in this question.

The support of standards

Standards live because they have supporters. The supporters do not have to be very loudspeaking as the history of OSIRIS shows. In the 70'ies OSIRIS was rather widespread as a social science analysis package, but SPSS[4] and SAS[5] gained momentum and OSIRIS was abandoned by the researchers. But many archives continued to utilise the OSIRIS documentational format, and still OSIRIS is used by many archives. Often the researcher that receives archive materials does not know that the SAS or SPSS setup actually is an automated product created from an OSIRIS codebook. When standards

depend upon supporters in order to get a strong standard you would naturally want strong supporters.

Two items express the need for commercial support: "18. A new format should be supported by the analysis software industry (e.g. SAS and SPSS)" and "19. A new documentation format should be supported by major document software and applications (Word, WordPerfect, WWW)". They both receive high agreement (mean 1.8). But does this mean that if we can not persuade SAS, SPSS, Word (Microsoft), WordPerfect (Novell) to support a new codebook format then we will have to give up? No, not in my opinion!

It would be nice with support from SAS and SPSS, but the history of data conversion shows[6] that the packages always almost do the job. That leaves you with problems concerning the character set and especially about missing data. Till now it has been much easier to write documentation conversion software that will reduce and format the documentation to the levels supported by SAS and SPSS. Another item indirectly addresses the commercial support: "21. A new documentation format will not be of any interest unless the data producers directly produce their documentation in this format". Here we are not only demanding software to follow the standards, but humans and maybe persons we know are asked to follow suit. The mean for item 21 drops to 2.5. Then it is getting really hot: "22. A new documentation format is only interesting if all archives abide by the new standard". The answer is close to indifferent (mean 2.7), but now we have moved from vendors to other people and finally we are trapped ourselves. The support of standards is a nice feature, and the less of one's own work that is involved, the nicer the feature gets.

Support from word processing companies will depend upon the new format. But if the new format is going to be directly connected to the formats of the Word Wide Web (HTML[7]) there is no doubt that the format is going to be supported[8] at least when the word processor is used as a viewer. However we have to be careful here. HTML is moving and changing, new versions and new features are being introduced. The safe ground to build upon is SGML where definitions can be made. Furthermore a codebook defined as a document type in SGML and marked up accordingly is very easily converted or reduced to any HTML level.

Labels in the codebook

New formats for codebooks or not, everything will not change overnight. We are going to continue to support analysis packages that can only handle limited amounts of text. But there is harsh disagreement that users or archives should be content with this level of information. "7. A documentation format with short labels for variables and values is sufficient for the user" (mean 4.0) and "8. A documentation format with short labels for variables and values is sufficient for the archive where a study is deposited" (mean 4.2).

When the question is directed specifically towards the length of the variable labels it is no great surprise that a longer label is preferred to the shorter: "14. Variable labels composed of 24 characters is sufficient" is slightly disagreeable (mean 3.4) whereas the next item is found slightly agreeable (mean 2.7): "15. Variable labels composed of 40 characters is sufficient"[9]. The question of course is what the labels are "sufficient" for? I interpret the results as specific for the labels, and not that the codebook documentation should consist of nothing more than the variable labels.

Oldies but goodies of the codebook

Information about the marginals is considered one of the main benefits of the codebook. "3. Codebooks should contain marginal frequencies that will enable the users to check the data they have received" is heavily agreed upon (mean 1.6). New fashions are not considered that important: "4. Codebooks should contain cross-tabulations so the user has more information about how to analyse the data" only makes it slightly above the threshold of indifference (mean 2.8).

Most of the items about the layout and printing of the codebook receives the same low level of attention. "9. The presentation of a printed codebook is very important" (mean 2.5); "24. There is little need for a printed codebook if a machine readable codebook exists" (mean 2.8); and "25. I prefer to browse data documentation in files on my own computer" (mean 2.6).

New possibilities of the codebook

If a new codebook format is to be developed we could be talking about transferring information, or we could use the opportunity to expand the current format with new possibilities. Some possibilities are mentioned amongst the assertions, and the highest score of all items is received by "20. A new documentation format should include the study description" (mean 1.4). Once again this demonstrates the methodological problem of having unlimited resources (points) when filling out the questionnaire. We must conclude from the questionnaire that there is a craving for including the study description in the codebook. From our practical lives we must conclude that we have to solve one thing at a time. Redesign of the codebook has to consider - but not necessarily to solve - the implementation of the study description.

I assume that most people are aware of the new possibilities but these are not considered very important for the codebook. The answer to "11. A documentation format should be able to incorporate pictures and sound" is indifferent (mean 2.9). The item on scanned images "23. I would be content to receive scanned images of the questionnaires" receives a slightly less favourable score, but that is due to poor verbalisation on my behalf. I believe that receiving only scanned images will not satisfy the user, but the combination of the text based codebook with the actual questionnaire as presented through scanned images will be of interest to most users.

Efforts of internationalisation have been mentioned at several meetings on documentation. One of the main obstacles in reading other languages than ones own. The introduction of English labels is not greeted with much enthusiasm: "6. English variable labels should be used with all studies regardless of the language of the original study" receives only a mean between agreement and indifference (mean 2.5). With the syndrome of unlimited resources you would have expected this to be easily agreed upon. Furthermore people from English speaking countries (UK, USA, Canada) seem to be only a fraction more inclined towards agreement.

The institutional questionnaire
The intention of the institutional questionnaire is to determine:

A) the number of datasets amongst the social science data archives.

B) the number of individual datasets (not held as a copy from another archive).

C) the distribution of used archive formats.

Datasets in social science archives
One of the big subjects has been to make a clear definition and to answer the question about the presumed unit of analysis "what is a study?". As archives would tend to see their importance depending on the magnitude of studies, and especially as compared with other archives, the study unit seemed to be of crucial importance. This led to some correspondence over the issue as it was argued that some studies consist of many datasets, have a complex scheme, and are distributed on several tapes, other studies have only few variables and a couple of hundred cases.

However I found that the introduction of a "true" unit of analysis could not succeed within the limited time period for returning the questionnaire. It should be noted that the individual questionnaire demanded only the person to make up his mind. In the institutional questionnaire more questions are asked based upon the unit of analysis. Even though the unit of analysis (the study) is not the same at all archives, it was more important that the individual archive had the easiest possibility of answering the questions; it was hard enough anyway. At each archive they had to do some kind of stocktaking and to place the studies within the categories demanded by the questionnaire.

Categorisations and fundamentals of archives
The archives that have answered the questionnaire are not many. 20 in numbers. One returned questionnaire was afterwards disregarded because both a staff member and the director of the archive had returned a questionnaire. 19 is now the basis.

The reason for some people to fill in the individual questionnaire as the only person from their archive, and yet not fill in the institutional questionnaire must be that the institutional questionnaire involved much more work in order to be filled out. Therefore I want to express my sincere thanks to the persons who took the time to fill out the time-consuming institutional questionnaire.

When preparing to discuss what kind of codebook documentation format should be used in the near future it is interesting to know whether an archive produces documentation. The answer of 7 archives to the question "Does your institution produce original machine readable documentation for studies in your archive?" was "No - we are only storing". Most of these archives are university data archives in North America, but the English national data archive was also a member of this group. If you do not produce documentation you will end up having all kinds of documentation formats. The interest of these archives in a new documentation format must be for the utilisation of the splendours of a new format, more than how a new format will act as a standard and solve some of our current documentational problems.

Of the 19 archives 13 archives have English as their native language.

Paper documentation vs. machine-readable documentation
First the respondent from the archive is asked about the number of studies with only paper documentation (Q1_1) secondly about the number of studies with some kind (any kind) of machine-readable documentation (Q1_2). The 19 archives share among them more than 27,000 studies, of these close to 11,000 have some kind of machine-readable documentation. The ratio of machine-readable documentation compared to all studies is thus 0.4. But this ratio hides great differences amongst the archives, some archives go as high as 0.9 others below 0.2 (the MR Ratio).

It is important to note, that the majority of studies at data archives do not have any machine readable documentation at all. Even when the machine readable text information is very limited (e.g. only describes variable labels and few categories) it can be used as input for software that automatically will convert the information to a new documentation format. This is a qualitative difference compared to having no machine readable text information. The studies without machine readable documentation will demand much more work to process to a new documentation format.

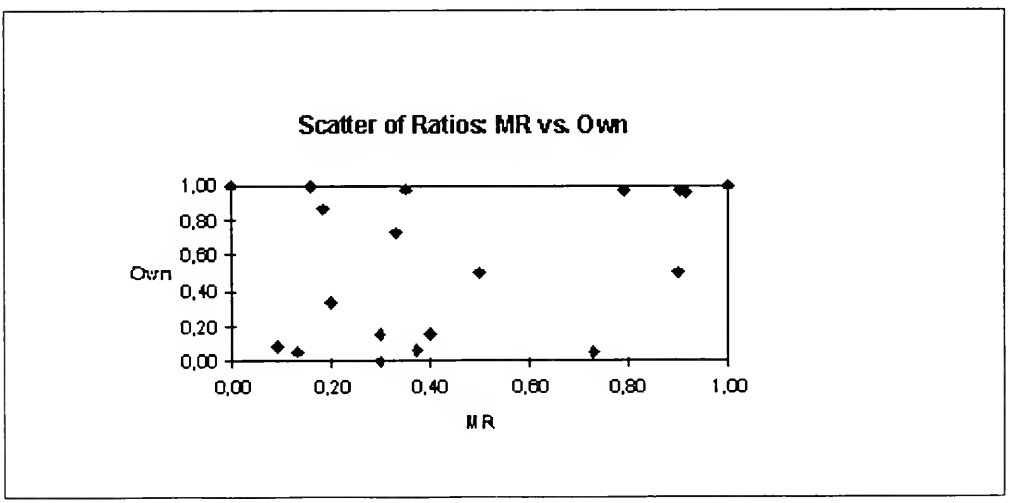
Own holdings vs. deposition from source archive
In Q2 the question is put about how many of the studies are actually stored and available from another archive (a source archive). The number of studies totals to more than 11,000. As many of these archives report their main source archive to be ICPSR, and as ICPSR is among the 19 answering archives we can decide to be bold and just subtract this from the total, leaving us with around 16,000 unique studies.

The ratio of own studies varies from 1.00 (all studies are own studies) to ratios close to zero (no studies that do not come from a source archive).

The scattergram shows the distribution between the ratio of own studies and the ratio of machine-readable studies in the figure below.

Types of MR
Then five questions about levels of machine readable documentation are asked. The respondent is asked to distribute the studies counted in Q1_2 having machine-readable documentation into five separate categories where the lowest rank is scanned images and the highest rank is a full codebook. The five categories are expected to sum to the total of machine-readable studies, but only half of the archives manage their stock data in such a way as to make the balance right. The sum of the studies in the five categories in Q3 totals to 12,627, whereas the total of machine readable studies in Q1_2 is 10,946.

Scanning as MR documentation
Scanning (Q3_1) of questionnaire pages as the highest level of machine-readable documentation is to close to 100 pct. only



found at the Amsterdam based Steinmetz archive. They have earlier shared their research and experience in scanning and made us aware of choosing TIFF-4 as the scanning format[10]. It is obvious that scanning is used at other archives both for security reasons as well as an easy deliverable - especially over Internet - but most studies will have some kind of higher level documentation as well.

One of the important things to remember when talking about scanning is that the scanned images should be referenced / pointed to / tagged from the character document that constitutes the codebook.

Format	Frequency
1. Scanning	505
2. Text	1943
3. Dict	4520
4. Dict+	3656
5. Dict + Codebook	2003
Total	12627

Plain text as MR documentation
After some editing[11] the distribution of question Q3_2 with the category "Unstructured and untagged text (text from OCR, questionnaire from WordPerfect, etc.)" ends with the following result:

At first it is surprising that the largest figure is given without any information about the format. On the other hand these formats are without importance because the formats are not directly related to the variables in the codebook or structured into elements of the variables. This category contains cases where the information is more like a stream of text for instance the non-edited result of OCR. When "ASCII" is mentioned this is not a reference to the irrelevant character format, "ASCII" here means "plain ASCII" which again means "no formatting information".

Dictionary as MR documentation
The category in Q3_3 "Machine readable dictionaries (information only about variable locations, labels, missing data in SAS, SPSS, OSIRIS or other format.)" means basic dictionary information on the variable level without information about the categories. In common database systems this is the dictionary information documenting the single fields of the database. However the introduction of information about missing data values and meanings is a social science product. A few formats are not widely used. For instance the Israeli archive mentions that they store their catalogue information in their ALEPH system, while the data is stored and documented in SAS.

Text Format	Frequency
ASCII	532
WordPerfect	299
Other	1112
Total	1943

Personally I do not find the differences in these systems interesting. As the dictionary level is defined in the question Q3_3 all these system will be very much alike. At the same time within a single product - for instance the much used SPSS system - the differences between different forms of SPSS can be as challenging to the user as receiving different products. SPSS can mean: Export files, System files (what platform), setups (what version) etc. 4,520 files are deliverable with the lowest level of machine readable variable level information.

Dictionary+ as MR documentation
This level is defined as the dictionary information plus information on values: Q3_4. "Machine readable codebook documentation (as 3. above but with the addition of explanation of the coding categories such as value labels in SPSS or user formats in SAS)". This is a restricted codebook format that does not have a lot of codebook levels and elements and does not support unlimited amounts of text.

Two interesting facts: DDMS is the system developed at Canadian Health and Welfare and the system is being used at different Canadian archives[12]. The NSD Stat is the format belonging to the statistical package developed by the Norwegian Archive[13].

Dictionary Format	Frequency
dBaseIV	20
OSIRIS	248
SAS	266
SPSS	3550
Other	436
Total	4520

Again the SPSS format is most often mentioned. And again my same views concerning formats between packages and within packages apply. Notice that the total number of studies archived drops when we demand value label information compared to the dict-level before.

Dictionary Codebook as MR documentation

In Q3_5 the level is defined as "Machine readable codebook documentation (as 4. above but including all questionnaire text and other information like in the OSIRIS format codebook)." This can give some difficulties if the respondent does not know the OSIRIS format.

DDMS	137
NSD Stat	100
OSIRIS	268
SAS	116
SPSS	2138
Other	897
Total	3656

Some are still mentioning SAS and SPSS, even though these packages cannot go above the level of dict+, but these packages are then combined with other text. Others mention different cataloguing systems and searching capabilities. All in all a conservative guess of the magnitude of fully documented studies can be as low as 1256 (DDMS plus OSIRIS). If we find this figure disappointingly low then on the positive side we can now remember that some archives have not answered the institutional questionnaire.

However these 2,003 studies plus the 3,656 studies belonging to the dict+ level are the studies with machine readable documentation available. These are the studies that the user will be able to analyse without consulting other material - except of cause the study description.

Dict Format	Frequency
DDMS	45
OSIRIS	1211
SAS	37
SPSS	37
Other	673
Total	2003

These 5,659 studies should optimistically be the studies that can be automatically converted to a new codebook format. If we are going to divide the task between us, I will personally choose to make the conversion of the OSIRIS codebooks because of their very simple format!

Conclusion

Future of the Codebook? Codebook of the Future!

We have seen that "what we want" is "everything".

"What we have" is pessimisticly close to nothing (1300 MR documented datasets).

We noticed that this high level of documentation should preferably be produced by "somebody else".

We know that only a small percentage of the studies are fully documented, and these can easily be converted to a new codebook format.

We know that more than half of the studies have no machine readable documentation at all.

We can conclude that we are facing a daunting big job.

However: If a new format can combine all (or almost all) our needs for structural elements in the documentation, the archivist will save time. I am very much looking forward to the work on a new format in the ICPSR "SGML Codebook Committee". We have a great opportunity in using the tools that are being offered.

Because the archives immediately will develop tools for processing the documentation for use in many surroundings like CD-ROM, Hypertext, Internet, etc. the possibilities of the new format will also be of interest to data producers. Laura Guy[14] talks about the archivist pleading with producers to ".. please .. document it properly?". With these modern add-ons there will be great benefits of doing proper machine readable documentation.

Paper presented at the IASSIST conference in Quebec City, May 1995

1 Report from SSD CESSDA seminar "Variable level documentation", Göteborg 1993.

2 The report and paper collection from the Grenoble meeting is not yet published.

3 The OSIRIS format is documented in OSIRIS III, Volume I, ISR 1973, Univ. of Michigan.

4 "SPSS for Windows, Release 6.0", 1993, Chicago, SPSS Inc.

5 SAS has meters of manuals, I'll mention 4 cm: "SAS Language: Reference, Verions 6", 1990, Cary, SAS Institute Inc.

6 Poster session at IASSIST 92 and "Converting Data" in DDA-Nyt 62, Summer 1992.

7 "Hyper Text Markup Language" is a document type definition (DTD) made in SGML (Standard Generalized Markup Language). HTML is used as the document format in WWW. A very usable SGML book is: "Practical SGML" by Eric van Herwijnen, 2. ed., 1994, Kluwer, Dordrecht.

8 In March 1995 Microsoft announced their HTML extension available for Word (but so far only for the US version 6.1).

9 As a curiosity I can mention that a cross tabulation of the two items (14 and 15) shows that some persons find 24 character labels agreeable but find 40 character labels disagreeable.

10 "Exchange of scanned documentation between social scientists and data archives: establishing an image file format and method of transfer".Repke de Vries and Cor van der Meer in IASSIST Quarterly Vol.16, number 1/2.

11 Apart from summarizing I have taken the liberty to evenly distribute figures connected to more than one format. If an archive gave the number 200 and mentioned the formats ASCII and WordPerfect both categories received 100.

12 William Bradley - mentioned in note 1 - is the leader of the group developing DDMS. It should be noted too that DDMS is a full codebook system, not a restricted system.

13 The documentation format is only mentioned at the Norwegian archive. But the use of the NSD Stat PC-package is much more widespread

14 "The Need for Revised Data Documentation Standards: New Solutions for Old Problems", Laura Guy in IASSIST Quarterly Vol. 17 Num. 3/4.

Appendix 1: The Questionnaires

In the following the questionnaires as e-mail to the listservers.

PART 1: DATA DOCUMENTATION PREFERENCES

Codebook Documentation of Social Science Data: an IASSIST Action Group

In this survey we are interested in your INDIVIDUAL opinionsabout improving data documentation and the formats used with data

documentation. The information obtained through this survey will become part of a report from an Action Group under the International Association of Social Science Information Service and Technology (IASSIST) about the current practices of social science data documentation and proposed standards for codebooks. Your thoughtful completion of this questionnaire is appreciated.

This first part is an individual questionnaire. If your institution stores or archives social science data we ask you to fill out the second part with information about your institution and archival format.

You are kindly invited to complete this questionnaire and return it to Karsten Boye Rasmussen, Dansk Data Arkiv, Islandsgade 10, DK-5000 Odense C., Denmark by mail, fax (+45 66113060) or e-mail (kb@dda.dk). If you are using e-mail please observe that you are not responding to the list server but directly to kb@dda.dk. Please return this questionnaire before the 15th of January 1995.

Q1. Below is a series of statements about data documentation. Please indicate for each statement the degree to which you agree or disagree with its content. Use the following five-point scale:

1 strongly agree
2 agree
3 indifferent, don't know
4 disagree
5 strongly disagree

__ There is no great need for standardization of codebooks.

__ A data user should be content with a study description and photocopies of relevant pages from the questionnaire.

__ Codebooks should contain marginal frequencies that will enable the users to check the data they have received.

__ Codebooks should contain cross-tabulations so the user has more information about how to analyze the data.

__ There is a great need for more structured information than is available in the OSIRIS codebook format.

__ English variable labels should be used with all studies regardless of the language of the original study.

__ A documentation format with short labels for variables and values is sufficient for the user.

__ A documentation format with short labels for variables and values is sufficient for the archive where a study is deposited.

__ The presentation of a printed codebook is very important.

__ Let us stick with commercially supported formats for social science data (e.g. SAS and SPSS).

__ A documentation format should be able to incorporate pictures and sound.

__ Missing data should always be coded as numeric values.

__ Coding of data fields using alphabetical and special characters should be discouraged.

__ Variable labels composed of 24 characters is sufficient.

__ Variable labels composed of 40 characters is sufficient.

__ Changing to a new documentation format would be very difficult to implement at our institution.

__ A new documentation format should be a specialized implementation of a general document format (e.g. SGML).

__ A new format should be supported by the analysis software industry (e.g. SAS and SPSS).

__ A new documentation format should be supported by major document software and applications (Word, WordPerfect, WWW)

__ A new documentation format should include the study description.

__ A new documentation format will not be of any interest unless the data producers directly produce their documentation in this format.

__ A new documentation format is only interesting if all archives abide by the new standard.

__ I would be content to receive scanned images of the questionnaires.

__ There is little need for a printed codebook if a machine readable codebook exists.

__ I prefer to browse data documentation in files on my own computer.

Q2. Your name : _____

Your position: _____

Institution: _____

Country: _____

Q3. Are you a member of IASSIST

( ) Yes
( ) No

Q4. Please comment further on any aspect of data documentation that is a concern to you.

PART 2: INSTITUTIONAL DATA DOCUMENTATION

Codebook Documentation of Social Science Data: an IASSIST Action Group

If your INSTITUTION stores or archives social science data we are interested in information about your institution and archival format.

You are kindly invited to complete this questionnaire and return it to Karsten Boye Rasmussen, Dansk Data Arkiv, Islandsgade 10, DK-5000 Odense C., Denmark by mail, fax (+45 66113060) or e-mail (kb@dda.dk). Please return this questionnaire before the 15th of January 1995.

This questionnaire is to be completed from an INSTITUTIONAL perspective. Individual opinions on social science data documentation are to be expressed in the first questionnaire. Several individuals from the same institution can fill out the individual questionnaire, but only one person needs to answer the institutional questionnaire.

The information obtained through this survey will become part of a report from an IASSIST Action Group about the current practices of social science data documentation and proposed standards for codebooks. Your thoughtful completion of this questionnaire is appreciated.

Your name: _____

Your position: _____

Name of your institution: _____

_____

Country: _____

Q1. We are interested in the variety of documentation that accompanies social science data and the number of studies with each type of documentation. Please indicate the total number of studies available with only paper documentation and those with some form of machine readable documentation. A study should only be counted once.

Number of studies archived with:

_____ 1. Only paper documentation
(no machine readable documentation)

_____ 2. Some form of machine readable documentation
(may also be available in print)

Q2. How many of your studies have you received from another archiving institution? Please specify number of studies:

_____ Studies from "source" data archive

Q3. Of the total number of studies with some form of machine readable documentation, please indicate how many studies are available in the following formats.

Please report a study only once
at the highest documentation level (1=low 5=high).

Number of studies in machine readable format consisting of:

_____ 1. Scanned images
Please specify the most commonly used format:

_____ 2. Unstructured and untagged text (text from OCR, questionnaire from WordPerfect, etc.)
Please specify the most commonly used format:

_____ 3. Machine readable dictionaries (information only about variable locations, labels, missing data in SAS, SPSS, OSIRIS or other format.)
Please specify the most commonly used format:

_____ 4. Machine readable codebook documentation (as 3. above but with the addition of explanation of the coding categories such as value labels in SPSS or user formats in SAS).
Please specify the most commonly used format:

_____ 5. Machine readable codebook documentation (as 4. above but including all questionnaire text and other information like in the OSIRIS format codebook).
Please specify the most commonly used format:

Q4. Does your institution produce original machine readable documentation for studies in your archive?

( ) no - we are only storing
( ) yes

If "yes" please describe the process and format used for preparing machine readable documentation.

Q5. Does your institution have a policy about the format used for documentation at the variable level?

( ) no
( ) yes

If "yes" please describe the format used for describing variables.

Appendix 2: Distribution and mean of the individual questionnaire

	strongly agree	agree	indiffe -rent, don't know	dis-agree	strongly disagree	Mean	N non-missing
1. There is no great need for standardization of codebooks	0	4	3	18	24	4.2	49
2. A data user should be content with a study description and photocopies of relevant pages from the questionnaire	1	3	2	14	30	4.3	50
3. Codebooks should contain marginal frequencies that will enable the users to check the data they have received	23	25	1	1	0	1.6	50
4. Codebooks should contain cross-tabulations so the user has more information about how to analyze the data	4	17	16	8	5	2.8	50
5. There is a great need for more structured information than is available in the OSIRIS codebook format	6	12	23	5	1	2.6	47
6. English variable labels should be used with all studies regardless of the language of the original study	5	25	10	6	4	2.5	50
7. A documentation format with short labels for variables and values is sufficient for the user	2	4	4	22	18	4.0	50
8. A documentation format with short labels for variables and values is sufficient for the archive where a study is deposited	2	3	4	14	27	4.2	50
9. The presentation of a printed codebook is very important	10	21	5	10	4	2.5	50
10. Let us stick with commercially supported formats for social science data (e.g. SAS and SPSS)	11	12	13	6	6	2.6	48
11. A documentation format should be able to incorporate pictures and sound	5	9	22	10	4	2.9	50
12. Missing data should always be coded as numeric values	15	12	14	3	6	2.4	50
13. Coding of datafields using alphabetical and special characters should be discouraged	18	11	8	9	4	2.4	50
14. Variable labels composed of 24 characters is sufficient	1	11	12	15	10	3.4	49
15. Variable labels composed of 40 characters is sufficient	7	16	12	9	5	2.7	49
16. Changing to a new documentation format would be very difficult to implement at our institution	6	4	17	17	5	3.2	49
17. A new documentation format should be a specialized implementation of a general document format (e.g. SGML)	12	17	14	4	1	2.2	48
18. A new format should be supported by the analysis software industry (e.g.SASandSPSS)	19	23	6	2	0	1.8	50
19. A new documentation format should be supported by major document software and applications (Word, WordPerfect, WWW)	18	25	6	1	0	1.8	50
20. A new documentation format should include the study description	29	19	2	0	0	1.4	50
21. A new documentation format will not be of any interest unless the data producers directly produce their documentation in this format	12	14	9	14	1	2.5	50
22. A new documentation format is only interesting if all archives abide by the new standard	7	14	16	11	2	2.7	50
23 I would be content to receive scanned images of the questionnaires	3	10	10	16	11	3.4	50
24. There is little need for a printed codebook if a machine readable codebook exists	9	17	1	17	6	2.8	50
25. I prefer to browse data documentation in files on my own computer	8	16	10	12	2	2.6	48

Tackling ICPSR Online Codebooks With Success

by Jackie Shieh [1]
ESRC Data Archive
University of Essex,

INTRODUCTION

The focus of cataloging the Inter-university Consortium for Political and Social Research (ICPSR) online codebooks is to provide users in a timely fashion adequate bibliographic information on VIRGO, the University of Virginia Library's computerized library system. Many catalogers today are cataloging materials that cannot be held in hand. Gathering bibliographic information for electronic formats can be a bewildering and monstrous experience. The author shares her experience on how the fear of working with computer files was reduced to a minimum with the help of the computer support department, and the sense of triumph and accomplishment she felt when patrons successfully retrieved what they needed through the online catalog!

BACKGROUND

The ICPSR is one of the world's leading repositories and data dissemination organizations for machine-readable social sciences data. ICPSR receives, processes, and distributes machine-readable data on subject matters covering over 130 countries. The content of the ICPSR archive extends across the spectrum of economic, sociological, historical, organizational, social, psychological, and political concerns. ICPSR was founded in 1962 as a partnership between the Survey Research Center at the University of Michigan and 21 universities in the United States. Currently, membership extends to over 370 colleges and universities world-wide. The University of Virginia (UVa) faculty and graduate students (undergraduates with the permission of an instructor) may order ICPSR data free of charge by contacting the Official Representative in the Social Science Data Center.

There are currently over 350 studies available at UVa. Most of the studies consist of more than one dataset, and often include a study description and a codebook. The codebooks which are in print or machine-readable format describe the location of the variables in the data record.

VIRGO, a NOTIS-based online catalog provides online access to the Library's holdings through keyword, author, title, subject, and call number searches. It also offers online access to: nine periodical indexes published by the H. W. Wilson Company; CURRENT CONTENTS (an index to recent issues of over 6500 scholarly journals); ABI/INFORM (provides citations and abstracts from business and management journals); and NEWSPAPER ABSTRACTS (indexes and abstracts articles in 28 major newspapers).

CATALOGING PROJECT FOR ONLINE CODEBOOKS

Although, Alderman Library began a cataloging project for paper codebooks in the summer of 1994, the project of cataloging the machine-readable codebooks did not take place until January 1995 with the arrival of the Original Cataloger for Electronic Resources. Before the project began for the machine-readable format, there were several issues that needed to be considered — 1) the number of titles incoming and in backlog, 2) the status of acquisition, 3) the procedure of retrieving bibliographic information, 4) the cataloging procedure, namely the procedure from getting actual text file to OCLC MARC format, and 5) the limitation for access of the materials.

Currently, the online codebooks reside in the machine named Maggie under the directory, /archive/public/icpsr at the University's Information Technology and Communication (ITC)[2]. They are accessible via all networks at the University. The files are sub-organized by the ICPSR series number from 0001-9999. Each individual series contains at least two basic files —the codebook and statistical data files.

To catalog a codebook, the following information needs to be obtained —the actual title, statement of responsibility, edition, file characteristics, physical description, series information, publication or distributor, special note and terms of availability, etc. The chief sources for the bibliographic description are taken from the title screen and table of contents.

Take for example, *CENSUS OF POPULATION, 1910. UNITED STATES: PUBLIC USE SAMPLE* (ICPSR ; no. 9166). To obtain the title proper, one can either pull up the title screen of codebook 9166 online from the sub-sub-directory of series number 9166 (*See* Figures 1 and 2), or consult the paper format reference book, ICPSR's *Guide to Resources and Services, 1994-1995* . Some of the series only have online codebooks, while others have both online and print versions. The Social Science Data Center at the University does not have the complete collection of neither online nor print version. Series titles are added as the faculty place subscription orders which result in the increase of the database, thus the ICPSR is a growing collection.

Once the bibliographic information is recorded, the title is searched against VIRGO and the national bibliographic utility, OCLC. If a record is found in VIRGO for paper

format (even if it is a different edition or release), the original cataloging for the computer file is created using the DER (derived) feature from the existing bibliographic record. If no VIRGO record is available and an OCLC record is found for a different format or edition, the OCLC printout is used to create an original record on VIRGO. All original cataloging for machine readable datafile are created locally on VIRGO. Currently, tape loading of machine-readable MARC records without field 856 from VIRGO to OCLC is not available.

FILTER PROGRAM USING PERL

The initial stage of the ICPSR project required the gathering of bibliographic descriptions on individual series numbers from the ICPSR directory. That task seemed far more cumbersome than the creation of the MARC record. It became even more overwhelming when it was discovered that new titles are checked-in and constantly added to the database. Therefore, in order to manage the existing and incoming codebook files tracking uncataloged materials became the priority.

Jeff Herrin, of the Library System Office wrote and explored a filter program using PERL (<italics>*See* <end italics> Figure 3) to read all current files in the directory of /archive/public/icpsr, from file 0001 to 9999. PERL copied the first 70 lines of text files, which corresponded to '*.codebk*', then copied the readings into an output file. The program would also generated an e-mail message notifying me even if there was no successful hit after the run. The first time PERL was run we had little difficulty. A few minor adjustments were made, especially when two codebooks for the same series number were present. For instance, series 8166 had two online codebooks, i8166.codebk and i8166.codebk01.

When examining the output of each file, I found that it did contain sufficient information for creating a cataloging record— the size of the file, title proper, author, publisher and publication date information (<italics>*See*<end italics> Figure 4). Although, the filter program was a success, the Social Science Data Center cautioned us that incoming materials are being deposited to their appropriate subdirectories regularly. In response to this, a cron daemon which runs the filter program automatically is scheduled every other month[3].

The VIRGO template (<italics>*See*<end italics> Figure 5) for ICPSR online materials was created to facilitate the cataloging process. Faculty, students and staff from the University can retrieve the information as soon as the bibliographic record is created.

CONCLUSION

After implementing the PERL filter program and creating the VIRGO template (similar to constant data on OCLC), cataloging online ICPSR series became more manageable than previously perceived. Yet, tapeloading from VIRGO to OCLC still presents a challenge for the Library. OCLC currently does not allow tapeloading on machine-readable original cataloging. Since the Library is committed to information and resource sharing, it means that contributing these bibliographic records requires additional manual work. Each record must be re-created for OCLC holdings. The Library is looking forward to the day when cataloging records for machine-readable format can be transferred

Figure 1

```
archive/public/icpsr% dir
drwxr-xr-x    4   3180  public  8192  Nov 23 09:07 0001-5999/
drwxr-xr-x   14   3180  public  8192  Nov 23 09:07 6000-6999/
drwxr-xr-x    9   3180  public  8192  Nov 23 09:07 7000-7250/
drwxr-xr-x        3180  public  8192  Nov 23 09:08 7251-7500/
drwxr-xr-x    7   3180  public  8192  Nov 23 09:08 7501-7750/
drwxr-xr-x   10   3180  public  8192  Nov 23 09:09 7751-7999/
drwxr-xr-x   11   3180  public  8192  Nov 23 09:09 8000-8250/
drwxr-xr-x   15   3180  public  8192  Nov 23 09:09 8251-8500/
drwxr-xr-x    5   3180  public  8192  Nov 23 09:09 8501-8750/
drwxr-xr-x    7   3180  public  8192  Nov 23 09:10 8751-8999/
drwxr-xr-x    9   3180  public  8192  Nov 23 09:10 9000-9250/
drwxr-xr-x   12   3180  public  8192  Nov 23 09:10 9251-9500/
drwxr-xr-x   12   3180  public  8192  Nov 23 09:10 9501-9750/
drwxr-xr-x   18   3180  public  8192  Nov 23 09:10 9751-9999/
```

Figure 2

```
archive/public/icpsr/9000-9250/9166% dir
-rw-r--r--  1  3180  public  408240    Nov 28 11:22  i9166.codebk
-rw-r--r--  1  3180  public  50965936  Nov 21 11:36  i9166.odata 1
```

Figure 3

```
#!/usr/bin/perl

$user="userid@virginia.edu";
$base="/lv2/users/userid/icpsr";
$tmpfile="$base/invent";
$pages="$base/pages";
$lastscan="$base/lastscan";

open(NEW, "find /archive/public/icpsr/*/* -name *codebk* -newer $lastscan -type
f -printl");
open(TMP,"> $tmpfile") ll die "Can't open tmp file!\n";
print TMP "Here's the new ICPSR codebooks:\n\n";

while($file=<NEW>) {
      print TMP $file;
      chop $file;
      $number = $file;
      $number =~ s/.*i([0-9]*)\.*cod.*/$1/;
      $edition = $file;
      $edition =~ s/.*\.codebk(.*)/$1;
      $size = -s $file;
      open(CODEPG,"> $pages/$number.$edition") ll die "Fail to open!\n";
      printf CODEPG "Size: %s bytes\n", $size;
      chop $file;
      $number = $file;
      $number =~ s/.*i([0-9]*)\.*cod.*/$1/;
      $edition = $file;
      $edition =~ s/.*\.codebk(.*)/$1;
      $size = -s $file;
      open(CODEPG,"> $pages/$number.$edition") ll die "Fail to open!\n";
      printf CODEPG "Size: %s bytes\n", $size;
      open(CODEBK,$file);
      while(<CODEBK>) {
            print CODEPG;
            last if $. > 80;
      }
      close(CODEBK);
      close(CODEPG);
}
`touch $lastscan`;
printf TMP "\nThe first pages are in $pages\n";
close(TMP);
close(NEW);
system("mailx -s \"New ICPSR items\" $user < $tmpfile");
system("rm -f $tmpfile");
```

figure 4

Size: 408240 bytes
1

CENSUS OF POPULATION, 1910 ^MUNITED STATESY:

PUBLIC USE SAMPLE

(ICPSR 9166)

Principal Investigator

Samuel H. Preston
University of Pennsylvania

First ICPSR Edition
Spring, 1989

Inter-university Consortium for
Political and Social Research
P.O. Box 1248
Ann Arbor, Michigan 48106

1

BIBLIOGRAPHIC CITATION, ACKNOWLEDGMENT OF ASSISTANCE
AND DATA DISCLAIMER

All manuscripts utilizing data made available through the Consortium should acknowledge that fact as well as identify the original collector of the data. In order to get such source acknowledgment listed in social science bibliographic utilities, it is necessary to present them in the form of a footnote or a reference. The bibliographic citation for this data collection is:

Preston, Samuel H. CENSUS OF POPULATION, 1910 ^MUNITED STATESY: PUBLIC USE SAMPLE ^Mcomputer fileY. Philadelphia, PA.: University of Pennsylvania. Population Studies Center, 1989 ^MproducerY. Ann Arbor, MI.: Inter-university

(END)

Figure 5

```
UL- ALK8984 FMT D RT m BL m DT 01/04/95 R/DT 01/25/95 STAT nn E/L  DCF a D/S D
 SRC d PLACE miu LANG eng MOD  T/AUD  D/CODE ? S/STAT ? DT/1 ???? DT/2
 DF/TYP d MACH  FREQ  REG  GOVT

 040:  :  a VA@  c VA@
 049:  :  a VA@@
 090/1:  :  a H62  b .I25 no.
 100:1 :  a <author>
 245:10:  a <title>
 256:  :  a Computer data (1 file : ca.  kilobytes).
 260:  :  a  Ann Arbor, Mich. :  b Inter-university Consortium for Political and Social Research,  c
<year>.
 490/1:1 :  a ICPSR ;
 500/1:  :  a Codebook to accompany related data tape.
 516/2:  :  a Text.
 516/3:  :  a <Numeric (Summary statistics).>
 520/4:  :  a <Optional.>
 500/5:  :  a <Also available in paper format.>
 580/6:  :  a Issued also in paper format, titled:
 537:  :  a Hard copy documentation (year) transformed into machine-readable text utilizing
Optical Character Recognition (OCR) Scanning, date.
 650/1: 0:  a <subject>
 700/1:10:  a <personal author>
 710/2:21:  a Inter-university Consortium for Political and Social Research.
 710/3:21:  a <corporate author>
 830/1: 0:  a ICPSR (Series) ;  v
 856/2:7 :  m Social Science Data Cender and ICPSR services, (804) 982-2630  u gopher://
gopher.lib.virginia.edu:70/11/socsci/icpsr  2 gopher.
```

successfully via either INTERNET FTP or tapeloading. The less editing required on one record, the more reliable the information remains.

1. Paper presented at IASSIST 95 Quebec City, Quebec. Jackie Shieh is Original Cataloger for Electronic Resources at Alderman Library, University of Virginia Library, e-mail ejs7y@Virginia.edu

2 The ITC is the equivalent of Computer Center in other institutions.

3. In UNIX, the cron daemon runs shell commands at specified dates and times. Regularly scheduled commands can be specified according to instructions contained in the crontab files. The cron daemon examines crontab files and at command files only when the cron daemon is initialized.

4. Tapeloading is available for the Internet Cataloging Project participating libraries under certain guidelines, Erik Jul's *Building a Catalog of Internet-Accessible Materials: Project Overview*. URL:http://www.oclc.org/oclc/man/catproj/overview.htm.

INTERNATIONAL ASSOCIATION FOR SOCIAL SCIENCE INFORMATION SERVICE AND TECHNOLOGY

• • • •

ASSOCIATION INTERNATIONALE POUR LES SERVICES ET TECHNIQUES D'INFORMATION EN SCIENCES SOCIALES

Membership form

The International Association for Social Science Information Services and Technology (**IASSIST**) is an international association of individuals who are engaged in the acquistion, processing, maintenance, and distribution of machine readable text and/or numeric social science data. The membership includes information system specialists, data base librarians or administrators, archivists, researchers, programmers, and managers. Their range of interests encompases hard copy as well as machine readable data.

Paid-up members enjoy voting rights and receive the **IASSIST QUARTERLY**. They also benefit from reduced fees for attendance at regional and international conferences sponsored by **IASSIST**.

Membership fees are:

Regular Membership. $40.00 per calendar year.

Student Membership: $20.00 per calendar year.

Institutional subcriptions to the quarterly are available, but do not confer voting rights or other membership benefits.

Institutional Subcription:

$70.00 per calendar year (includes one volume of the Quarterly)

I would like to become a member of IASSIST. Please see my choice below:

- ☐ $40 Regular Membership
- ☐ $20 Student Membership
- ☐ $70 Institutional Membership

My primary Interests are:

- ☐ Archive Services/Administration
- ☐ Data Processing
- ☐ Data Management
- ☐ Research Applications
- ☐ Other (specify)_____

Please make checks payable to IASSIST and Mail to :
Mr. Marty Pawlocki
Treasurer, IASSIST
% 303 GSLIS Building,
Social Science Data Archives, University of California, 405 Hilgard Avenue, Los Angeles, CA 90024-1484

Name / title

Institutional Affiliation

Mailing Address

City

Country / zip/ postal code / phone